Machine vision (computer vision) is an important area of research dealing with the processing and analyzing of image information. There are numerous potential areas of application for machine vision, including industrial inspection, robotics, vehicle guidance, medical image analysis, document analysis, remote sensing, visual surveillance, human-computer interfaces, identity authentication, multimedia, and scene modeling for virtual reality. Possibilities for processing, analyzing and transmitting image/media information are continuously increasing due to the increased performance of computers and speed of telecommunications. Various new applications for digital media processing and services are emerging.

The mission of the group is to investigate techniques and applications of machine vision and digital media processing. Methods and systems for important application areas based on original algorithmic and engineering solutions are to be developed. In order to reach the goals it is important to carry out long-term fundamental research and to have comprehensive system-level know-how of the problem areas to be solved. New, powerful and sufficiently generic algorithms and other engineering solutions have to be developed, experimental systems for chosen applications have to be built, and extensive performance tests in the laboratory and in the real application environment have to be carried out. The group believes that this kind of system level approach which includes long-term research in core areas, and close interaction with application scenarios is needed to make good progress and to ensure efficient technology transfer and adaptation. This approach is not common among research groups in this field.

The goal of the group is to be an internationally highly ranked and respected research unit, the activities of which can be seen from its high quality publications and doctoral dissertations, internationalism, ability to conduct research according to the changing needs of applications, and efficient technology transfer of research results from the laboratory to practical applications. In 1997, the group was ranked by the University of Oulu among its internal research units of excellence.

In the past, the main research area of the group has been industrial machine vision, in which it has reached an internationally highly respected position. Recently, the group has widened its activities to include digital media processing and services. The objective is to rapidly raise the quality of research to world-class level in this area as well.

The topics of the long term research of the group are chosen to serve the needs of the principal application areas. The current core areas of research are: 1) color and texture based scene analysis; with applications in surface inspection, surveillance, medical image analysis, and media processing, 2) image sequence processing and transmission; with applications in industrial automation, visual surveillance, wireless image communications, media processing, and medical image analysis, 3) media processing and services; document analysis, intelligent retrieval of documents, images and videos from databases and their bandwidth saving transmission in networks in a multimedia context, computer telephony integration, mobile multimedia services, 4) medical signal and image processing; with applications in aerobic fitness determination from heart rate measurements, diagnosis of skin diseases, and magnetic resonance imaging (MRI), and 5) sensor based control of intelligent mobile robots.
The group has continued to carry out long term research in its five principal research areas, achieving many significant results in all areas.

The "Cooperative research on computer vision” project with the University of Maryland (USA), which is mainly funded by the Academy of Finland and Tekes, was continued. In this three-year project, new systems and algorithms for document analysis, media processing and image analysis are being investigated. An industrial Esprit project "Color and texture inspection equipment” funded by the European Commission is in its second year of progress. The members of the group have played central roles in this project dealing with methods, systems and applications of automated visual inspection. Prof. Olli Silvén is the technical coordination manager of the whole project. The group is a primary node of the European Computer Vision Network (ECVnet) and has participated in its activities by, for example, organizing a Technology Transfer Workshop in Machine Vision in Lappeenranta, Finland, in June 1997.

The scientific progress in the five research areas (1-5) is described in the following paragraphs:

1. Color and texture analysis belong to the important generic research areas of machine vision. A methodology for measuring very small color differences with a RGB color camera, based on physical color models and online camera calibration, has been developed. The methods were tested in the inspection of colored paper and food samples, both in our specially designed color laboratory and in industrial environment. New research on a smart camera with color correction based on scene content was started. Nonparametric approaches to texture analysis using spatial operators have been developed. Excellent results in texture classification and unsupervised texture segmentation were achieved. Different density estimation methods have been studied in collaboration with the University of Maryland, in association with multiband segmentation and classification of satellite imagery.

2. Motion analysis based on the use of image sequences is also a generic area of machine vision with many applications. Methodology and new algorithms for carrying out precise measurements using monocular image acquisition have been developed. Novel error concealment methods for error resilient decoding of digital video were created that exploit estimates of actual 6-dof motion of the camera. The tests have been carried out using rigid scenes. Solutions for supporting mobile video transmission over wireless and ATM-networks have been developed and demonstrated. The very low latency handovers have been achieved that are necessary in real-time applications.

3. An intelligent document image retrieval system has been developed in cooperation with the University of Maryland. It was also generalized and integrated for retrieval of ordinary scene images. A distributed environment for testing document analysis methods and systems utilizing distributed network computing in the media domain has been developed, as well as a system for benchmarking document analysis algorithms. New algorithms for different stages of the document analysis process were developed and a new concept of “hyperdocument” was proposed. A new intelligent document and media preprocessing system was devel-
4. An intelligent signal analysis method for approximating aerobic fitness from demographic and heart rate variables using an artificial neural network has been developed. An image analysis system for psoriasis area assessment has been developed, as well as an algorithm to register lesions in successive skin images. A fast registration method for operative and preoperative MR images has been developed, as well as novel concepts for MRI guided minimally invasive surgery.

5. A control architecture for basing the actions of a mobile robot has been developed in cooperation with the Electrotechnical Laboratory (Tsukuba, Japan). The architecture is behavior-based, capable of reactive task execution - of both reasoning actions based on task constraints and reacting quickly to unexpected events in the environment. An obstacle detection system using a light stripe identification based method has been developed. The method makes the structured light based detection system more robust and applicable to use outdoors as an aid for navigation.

EXPLOITATION OF RESULTS

The results of the research are being widely exploited in all the main research areas:

1. Inspection methods developed by the group have been applied to food sorting and wood grading in the Esprit project “Color and texture inspection equipment”, and used in the wood grading system developed for VTT Construction Technology. The use of color cameras from different manufacturers for accurate color measurements has been studied. A tool for interactive multiband image segmentation and classification has been developed for the Department of Geography of the University of Maryland.

2. A precise camera calibration method has been developed for the metallurgical vessel lining wear system of Spectra-Physics VisionTech. A Matlab-based calibration toolbox available for public use has been distributed via the Internet. The tool has attracted numerous users both in industry and academia. A traffic counter, as well as digital video transmission technology for a wireless multimedia system used for teleoperating mining machines have been developed for Elektrobit Inc. A 3D teeth measurement system has been developed for the Department of Dentistry of the University of Oulu, and is being duplicated at a foreign institution.

3. The group has played a central role in the planning of the Mobile City Oulu technology program, in which advanced services for next generation mobile communication systems will be developed by the University of Oulu, VTT Electronics and industry. The results of the group’s research in document processing and image sequence processing will be exploited. Cooperation with Nokia in the area of computer telephony integration has been started. The development of scalable mobile end-to-end services to several network and terminal types has been started in collaboration with VTT, several companies, the University of Maryland and NIST, USA.
4. The image analysis method was applied to quantify affected surface area of psoriasis from color photographs. The reliability of the image analysis was tested with a reference analysis by coloring manually the involved area in the images and the agreement was found to be high. The use of different texture features for the classifications of melanocytic lesions has been studied. A European Esprit project “Intraoperative real-time visualization and instrument tracking for MRI” applying our results in image sequence processing was under preparation. The proposal has been accepted and the project started in January 1998.

5. The control architecture developed by the group has been applied in playing simulated soccer. As the only Finnish team the group participated in the first RoboCup competition held in Nagoya, Japan. Although the main effort of the group was to show that the controlling system can be used to create independent players with certain skills which they could modify and tune according to prevailing situation, the group could also show that independent players could operate as a team.

**Future Goals**

Recently, our funding for media processing and services has substantially increased and the total external funding of the group for 1998 is over 8 million FIM. Due to this our future research can be divided into the following main areas: 1) machine vision and intelligent systems and 2) media processing and services. Our activities in these two areas are closely related and support each other. We believe that this will also be our greatest strength in media processing: our strong expertise in machine vision and intelligent systems combined with our rapidly growing expertise in multimedia and telecommunication networks provide excellent opportunities for creating various new types of services, e.g. for mobile communications and value added telecom systems.

In the machine vision and intelligent systems area, we plan to continue our long term research into color and texture analysis, image sequence analysis, intelligent signal analysis and intelligent robot control, creating in this way a basis for the development of new and more advanced applications in the future. A growing part of this research is directed to support our activities in media processing and services. In that area, our research will be focused on document and video analysis, content-based media retrieval, distributed media processing, computer telephony integration, and mobile multimedia services.
DOCTORAL THERSES


SELECTED PUBLICATIONS


Background and Mission

Machine vision (computer vision) is a continuously growing area of research dealing with processing and analyzing of image data. There are numerous potential areas of application for machine vision in various fields. Possibilities for processing, analyzing and transmitting image/media information is continuously increasing due to the increased performance of computers and speed of telecommunications. Various new applications for digital media processing and services are emerging.

The mission of the unit is to carry out leading edge long term research on machine vision and media processing technology, with an aim of bringing the results of the research close to applications.

The unit plans to carry out its mission by performing long term basic research in four core areas, and applied research on several important application problems. In order to be able to bring its research results near to applications, profound knowledge on key systems engineering technologies will be maintained and further developed.

The core areas of research are: image analysis, image sequence analysis, intelligent systems, and media processing and networks.

The research on image analysis concentrates on problems in texture analysis, color analysis and document image analysis. The focus of image sequence analysis is in tracking moving objects, like human beings, from video sequences, efficient transmission of video streams and 3-D machine vision. In intelligent systems, control systems for autonomous mobile robots and techniques for intelligent signal analysis are considered. Media processing and networks concentrate on media adaptation and on distributed processing and networks with their utilization in complete end-to-end systems.

The research in the above core areas provides the basis for original applied research in various areas. The areas of interest include: inspection and sorting, visual surveillance and video indexing, video coding for wireless channels, geometric camera calibration and industrial metrology, intelligent robot control, biomedical signal and image analysis, information retrieval, mobile information systems, computer telephony integration, and media telephony.

The unit cooperates with many international and domestic partners. In basic research it has joint projects with the University of Maryland (USA) and the National Institute of Standards and Technology (USA). In applied research, the unit plays central roles in two European Esprit projects and several joint projects funded by Tekes and industry.

The activities of the unit are lead by professors Matti Pietikäinen (director), Olli Silvén (associate director), Juha Röning and Jaakko Sauvola.

In 1998, the unit was chosen among the top 51 candidates, representing all fields of science, in the competition for the status of a national center of excellence. The international evaluators ranked the unit excellent and regarded it as a good candidate for obtaining the status of center of excellence.
Scientific Progress

Research on solutions needed for making a color camera based vision system work reliably in uncontrolled environments was in progress. A unique facial image database containing both image and spectral data was created. Methods for correcting color images taken under uncontrolled illumination were investigated. A method combining a physics-based approach and principal component analysis (PCA) provided very promising results in the correction of face images.

Research on nonparametric approaches to image texture analysis was continued. A method based on multidimensional distributions of signed gray level differences was developed, providing excellent performance in texture classification. An efficient method based on learning vector quantization was developed to reduce the multidimensional distributions into one dimensional histograms.

A Self-Organizing Map (SOM) based classification and user interface scheme has been developed primarily for surface inspection applications, but appears to be useful outside this field, for example in video classification. The approach supports the labeling of training data, simplifies retraining in the case of changing material or imaging conditions, provides an intuitive user interface, and is computationally attractive for real-time use.

A new method for document skew estimation without angle range restriction was developed. Document skew is a distortion that often occurs during scanning or copying of a document. The method is based on determination of the first eigenvector of the data covariance matrix. Experiments on a large set of various document images showed good accuracy and robustness.

A geometric camera calibration procedure for precise 3-D computer vision applications using circular control points has been further developed. Camera based coordinate measurements are typically made for relatively large scale objects. In our recent experiments, a precision greater than 5 microns was achieved and with fully digital image acquisition hardware a precision of one micron seems to be feasible. The results indicate that the camera based measurement technique is also applicable to smaller dimensions with no degradation in relative accuracy.

Methods for finding and tracking people from image sequences have been studied extensively. The performances of many state of the art approaches were experimentally evaluated. This prepared the way for tackling the more fundamental problem of finding people and interpreting their actions from sequences acquired by cameras in changing environmental conditions.

An error concealment method for motion compensated
block coded real-time video has been developed. The approach estimates the 3-D motion of the camera and hides the missing regions using previous images in the sequence. Good concealment results are achieved by properly scaling the correcting patches using the estimated focus of expansion and optical flow vectors. The method has been demonstrated with wireless teleoperation video sequences.

Control architecture for a mobile robot operating in a dynamic environment has been extended to manage multiple independent robots (agents) with an application in simulated soccer. The first generation of software agents traveling in the Internet and performing simple tasks was also demonstrated.

For 3-D environment modeling, an experimental system is being developed that combines 3-D information provided by structured lighting with the 2-D information provided by intensity (color) images. As an application, the creation of a model of a building for virtual reality applications has been considered.

Research on intelligent signal analysis methods for approximating aerobic fitness from demographic and heart rate variables has continued. The results obtained with a neural network approach have been improved using principal component analysis techniques. The results achieved by neural network have also been compared with those obtained by regression tree approximations.

Research on intelligent document image retrieval has continued. The content-based multimedia document image retrieval system developed by the unit is among the first systems in the world in its field that embeds active multimedia properties.

Research on scalable mobile end-to-end services for several network and terminal types has been in progress. Scalable processing algorithms and methods for digital media services over hybrid network architectures to exploit web information and multimedia documents have been developed.

**Exploitation of Results**

Inspection methods developed by the group have been applied to food sorting and wood grading in the Esprit project “Color and texture inspection equipment”, and used in the wood grading system developed for VTT Building Technology.

In the European Esprit project called “Intraoperative real-time visualization and instrument tracking for MRI” our results in image sequence analysis are being exploited. The project will demonstrate the feasibility of an interactive MRI intervention system, which features real-time visual feedback.

The image analysis method developed for assessing affected surface area in patients with psoriasis has been experimentally evaluated from digital photographs taken at the beginning and at the end of different anti-psoriasis treatments. The method offers a possibility to quantify the actual surface in patients with psoriasis, and will be an alternative for developing quality control when
evaluating different treatments.

The control architecture developed by the group has been applied in playing simulated soccer. The unit participated in the simulation league of the second world cup of robotic soccer, RoboCup’98, held in Paris in July. Our team finished among the 8 best teams of the 32 participants.

One of the mobile robots appeared in the “Homo Mobilis - Moving Man” exhibition at the Museum of Art and Design in Helsinki, Finland. The robot (a Nomad Super Scout from Nomadic Technologies) wandered in a living room of the future. This room demonstrated how man can move in the future while remaining on the couch. Our robot wandered there for 6 hours every day and expressed in Finnish its ideas about the robot of the future.

A large test database of document images, the ‘Oulu Document Database’ with ground-truth classifications and tools for content-based search, was created. The database is distributed via CD-ROM and is currently under second evaluation in several research groups.

A new distributed computer-telephony platform was developed that exploits media-based intelligence and object-oriented software agent technology. The research was commissioned by Nokia Telecommunications, where the results are now being used.

New media telephony concepts were developed for wireless office communications that use media understanding, distributed multimedia processing and wireless access. This work was performed for Nokia Wireless Business Communications.

**Future Goals**

The unit plans to continue its long term research in all core areas and to actively seek possibilities for exploitation of results. In 1998, the unit prepared its long term research plans for 2000-2005, developed its organization and management to meet the new challenges ahead and took steps to improve the publicity of its research. In the near future, the unit plans to hire more visiting researchers and support personnel and to further intensify its international activities, if funding can be arranged.

The enterprises of the Oulu region in the area of information technology have grown very rapidly in recent years. This has significantly increased their interest in our research, especially in the area of media processing. Partly due to this, the level of funding and the size of the unit has recently significantly increased. This has created many new opportunities and challenges for our research.

In order to keep its research well focused with this development and to improve its possibilities to obtain necessary funding from various sources, the unit is considering a split into two closely cooperating groups: one concentrating on machine vision and intelligent systems and the other on media processing and networks.
### Personnel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>doctors</td>
<td>9</td>
</tr>
<tr>
<td>graduate students</td>
<td>15</td>
</tr>
<tr>
<td>others</td>
<td>37</td>
</tr>
<tr>
<td>total</td>
<td>61</td>
</tr>
</tbody>
</table>

### External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>FI$\text{Ms}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland + Ministry of Education</td>
<td>2 022 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>3 217 000</td>
</tr>
<tr>
<td>Infotech + center of excell. (Univ. of Oulu)</td>
<td>433 000</td>
</tr>
<tr>
<td>domestic private sources</td>
<td>1 959 000</td>
</tr>
<tr>
<td>EU + other international sources</td>
<td>610 000</td>
</tr>
<tr>
<td>total</td>
<td>8 241 000</td>
</tr>
</tbody>
</table>

### Doctoral Theses


### Selected Publications

In 1999, the Machine Vision and Media Processing unit was split into two closely cooperating groups: one concentrating on machine vision and intelligent systems (MVIS) and the other one on multimedia, mobile computing and value adding services (MediaTeam Oulu). Both of them were selected for the new three year period to Infotech Oulu. The groups are described separately, but the following statistics and publication information concern the whole unit.

### Personnel

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>16</td>
</tr>
<tr>
<td>graduate students</td>
<td>21</td>
</tr>
<tr>
<td>others</td>
<td>48</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>85</td>
</tr>
</tbody>
</table>

### External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>FIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>1 472 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>1 413 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>4 718 000</td>
</tr>
<tr>
<td>Infotech + center of excell. (Univ. of Oulu)</td>
<td>406 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>3 029 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>518 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>11 556 000</strong></td>
</tr>
</tbody>
</table>

### Doctoral Theses


### Selected Publications


MACHINE VISION AND INTELLIGENT SYSTEMS

Professor Matti Pietikäinen and Professor Juha Röning,
Information Processing Laboratory and Computer Engineering Laboratory,
Department of Electrical Engineering, University of Oulu

Background and Mission

Machine vision is a continuously growing area of research dealing with processing and analyzing of image data. It plays a key role in the development of intelligent systems. The principal goal of research on intelligent systems is to develop machines that have human-like features, such as the ability to perceive, reason, move, and learn from experiences.

The mission of the MVIS group is to carry out leading edge long term research on machine vision and intelligent systems technology, with an aim of bringing the results of the research close to applications.

The MVIS group plans to carry out long term research in three core areas and apply the results of research to several important application problems. In order to be able to bring its research results near to applications, considerable in-depth know-how on key systems engineering technologies will be maintained and further developed. With this approach, the group aims to bridge the gap that exists between scientific research and industrial practice.

The core areas of research are:

- Image analysis
- Image sequence analysis
- Intelligent systems

The research on image analysis concentrates on problems in texture analysis, color analysis, and document image analysis. The focus of image sequence analysis is in detecting and tracking moving objects, like human beings, from video sequences, video indexing, video coding for wireless real-time transmission, industrial video camera based metrology, and interventional MRI. The research on intelligent systems concentrates on smart environments, intelligent mobile robots, neural network methodology for signal analysis, and analysis of biomedical ECG and EEG signals.
The research in the above core areas provides the basis for original applied research in various areas, including industrial automation, mobile communications, multimedia, intelligent interfaces and medical engineering.

The group cooperates with many international and domestic partners. In basic research it has close collaboration with the University of Maryland (USA). In applied research, the group has played central roles in European Esprit projects and several joint projects funded by Tekes and industry. The group is also active in the scientific community. For example, in June 1999 it organized a successful Infotech Oulu Workshop on Texture Analysis in Machine Vision.

The activities of the MVIS group are led by professors Matti Pietikäinen (Director), Juha Röning (Associate Director), Olli Silvén and Tapio Seppänen.

**Scientific Progress**

**Image analysis**

Image texture analysis has been a topic of intensive research for over three decades. During the past few years, the group has developed theoretically and computationally very simple nonparametric approaches to texture analysis based on Local Binary Patterns (LBP) and signed gray level differences. In 1999, LBP-type operators for gray scale and rotation invariant texture classification and multiscale versions of LBP were investigated. The use of our methods for classifying tilted 3D textures was also studied. All these methods outperformed the state-of-the-art approaches in extensive comparative studies that were carried out. The usefulness of the techniques, based on Fourier spectra for extracting perceptually significant information from textures, was also evaluated.

Illumination invariant techniques for color face image processing are being pursued. Color eigenfaces computed from the University of Oulu Physics-Based (UOPB) Face Database were used in a technique developed to compress and correct color shifts in frontal view of faces taken under uncalibrated camera conditions. Color is a primary cue in detecting faces or hands in video, but it is dependent on illumination. Properties of skin chromaticity in normalized color space are being exploited to extend skin detection even under changing lighting conditions. The UOPB Face Database was made available in CD-ROM and has been delivered to several academic and industrial research groups all over the world.

In the ongoing research on plant vitality measurement, the purpose is to find out whether the vitality of plants can be determined from their reflectance spectra in greenhouse conditions. The measurement instrument used is an imaging spectrograph, which allows the exact segmentation of the plant parts from the background. The challenges of the research are in determining the model for the correlation of the vitality and the spectrum, determining the specifications for the measurement device and coping with the varying illumination conditions.

In wood surface inspection, the use of new texture inspection methods is being investigated. The SOM (self-organizing map) based tool for visual inspection developed in our group is adapted for the wood inspection to relieve the training material management. A detailed description of a color-based wood inspection method and more general aspects in method development and evaluation were published in a dissertation.

In document image analysis, an extensive survey on page segmentation and zone classification methods for document layout analysis was performed in collaboration with the University of Maryland. Research on the development of an advanced document layout analysis system was begun. A methodology for automatic ground-truth generation is being developed in order to evaluate the skew-tolerance of page segmentation methods.
**Image sequence analysis**

Methods for detecting humans from image sequences have been developed based on two approaches. A computationally less demanding method is to use human skin detection from color images. The experiments have shown that this method has potential in surveillance applications. Another approach considered is motion detection based on successive frames. Some new motion detection and segmentation algorithms have been developed for tolerating camera movements and illumination changes. In human tracking, a method for modeling the background objects with a Gaussian mixture has been used, together with a Kalman filter based tracking algorithm.

Error resiliency features of the MPEG-4 codecs have been tested extensively. In wireless transmission systems, considerable improvements in video quality can be achieved with these features. In those situations where error correction cannot recover the lost image data, error concealment is applied. For that purpose, a new error concealment method utilizing the redundancy between the transmitted motion vectors has been developed.

In geometric camera calibration, a new algorithm for coplanar target structures has been developed. The main advantage of using coplanar targets is that precise 2-D objects are much easier to make than 3-D structures. In order to achieve accurate calibration results with 2-D targets, several images from different viewpoints must be captured. Traditionally, these images are processed simultaneously, unlike in the new algorithm, where each image is processed separately. With this technique the computational complexity of the calibration procedure can be reduced significantly.

In 3-D measurements, a low cost single camera based system for modeling the profile of hot steel making vessels was built. In the experiments, the system was capable of measuring targets in a temperature of 800-900 degrees. Modifications to the lighting solution are needed for measuring in higher temperatures. Also, other application areas for the single camera based measurement technique are being investigated.

**Intelligent systems**

Research on developing control architecture for a mobile robot operating in a dynamic environment has continued. The work has proceeded to implementing a teleoperation system which co-operates with the control architecture. The low-level reactive control architecture prevents the user making simple errors like colliding with obstacles. The user interface is a Java applet running in the popular web browser, so it can be used in any place having an Internet connection (with sufficient bandwidth) to the premises of the robot. The next step is to perform experiments with the mobile robot in a local residential home.

The control architecture developed by the group has been applied to playing simulated soccer. Our RoboCup’98 team acted as a qualifying competitor for the RoboCup’99. Our student group participated in the simulation league of the third world cup of robotic soccer, RoboCup’99, held in Stockholm in June.

For 3-D environment modeling, an experimental system is being developed that combines 3-D information provided by structured lighting with the 2-D information provided by intensity (color) images. During this reporting period a more robust and integrated version of the system has been constructed.

Research on neural network methodology for signal analysis has focused on spatially and temporally adaptive learning applications and techniques. A static and deterministic application cannot cope with all the temporally changing situations in, for example, industrial environments or notably dynamic environments, such as the human physiological system.
The data used includes a huge amount of registered bio-signals, such as heart rate, temperature, blood pressure as well as subjective diary recordings about the person’s overall subjective health status. One goal is to find explanations for phenomena visible in bio-signals from diary recordings. A software application has been developed for an experimental learning and adaptive system to recognize, for example, stress situations in a person’s daily life.

New methods for the monitoring of human physical signals in everyday life have been developed, but in order to get more reliable results, longer measurement periods with different life situations (stress, illness, holidays, etc.) are needed. In the pre-processing state, variables obtained with spectral analysis cannot be utilized with this data for the same reason. More data from different populations is being gathered continuously.

Research into ECG signals has been conducted. The main goal in the project is to study heart rate variability of patients and normal people in order to develop methods for automatic diagnosis and prognosis for clinical use. The activity is mainly performed in the Oulu University Hospital, but a network of industrial companies and universities is also involved. During 1999 two Ph.D. degrees in medicine were achieved and many journal papers were published and accepted for publication. An example of more technically oriented papers is given in the list of publications.

New research on EEG signals has been launched in 1999. The goal of this study is to develop methods for automatic determination of deepness of anaesthetics in clinical operations. This goal is approached by performing morphological analysis on evoked potentials and phase synchronization of sub cortical cell populations. This research is conducted at the Oulu University Hospital in cooperation with other Finnish universities and industrial companies.

**Exploitation of Results**

The results of our research were applied to real-word problems in many projects, often in collaboration with industrial and other partners. Some examples of exploitation are described below.

Inspection methods developed by the group have been applied to food sorting and wood grading in the Esprit project entitled “Color and texture inspection equipment” that ended in 1999, and used in the wood grading system developed for VTT Building Technology.

The European Esprit project called “Intraoperative real-time visualization and instrument tracking for MRI” (IRVIT) demonstrated the feasibility of an interactive MRI intervention system, which features real-time visual feedback. The IRVIT project ended in 1999, but a follow-up project proposal (AMIT) targeted on clinical trials was accepted by the European Commission.

The use of a mobile robot for domestic help has been studied. A teleoperated robot serves as the remote eyes of the elderly and those who take care of them. During the reporting period, the main task has been to develop teleoperation capabilities for the robot. The first successful usability tests were performed. The goal is to create new product and service concepts for helping the elderly and disabled in their everyday lives. It aims at decreasing the need for institutional care by developing a communication system that enables the elderly to keep in contact with their relatives and friends and to run errands from their homes.

**Future Goals**

The MVIS group will continue to strengthen its long term research. We plan to be increasingly active in the international research community, for example by submitting more articles to scientific journals, arranging workshops, intensifying international cooperation and researcher exchange, and delivering our research results and test material via the Internet. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects.
Background and Mission

MediaTeam, in its present form, came into being in 1999 when the thriving Machine Vision and Media Processing unit of the Information Processing and Computer Engineering Laboratory in the University of Oulu was divided into two closely collaborating but independent research groups, MVIS and MediaTeam, in order to facilitate efficient research in terms of focus and resources.

The mission of MediaTeam is to carry out leading edge long-term research, which produces new scientific knowledge and novel technological solutions to distributed multimedia and communications.

MediaTeam carries out research in the fields of multimedia, mobile computing and value adding services, and is positioned on the overlap between information technology and telecommunications. Our interdisciplinary approach to research manifests itself in our connections with various departments and universities in different fields. Another example of our unconventional approach is our personnel, which consists of engineers and natural scientists as well as linguists.

Three core areas of research have been identified: multimedia understanding, multimedia transmission, and multimedia services. **Multimedia understanding** is concerned with a variety of issues ranging from media types and formats to the refinement of analysis information. **Multimedia transmission** focuses on transmission-related topics such as content-based transmission, networked computing, and the quality of the provided services. Our final core area of research, **Multimedia services**, combines the results obtained from the two previously mentioned entities in order to produce efficient, scalable applications and services for the benefit of distributed, mobile environments.

MediaTeam comprises of five competence teams, corresponding to the five themes which have been identified as pertinently focusing our research:

- Content-based Multimedia Retrieval, CMR
- Computer-Telephone Integration, CTI
- Wireless IP and Media Telephony, MET
- Mobile Information Systems, MIS
- Multimedia Signal Processing, MSP

The activities of the MediaTeam research group are led by Professor Jaakko Sauvola (Director), Dr. Timo Ojala (Associate Director) and Professor Tapio Seppälä (Scientific Director).
Scientific Progress

The main emphasis has been put on the further development of service platforms that provide the framework for ongoing and future algorithm and application level research. The platforms are as follows:

- **CMRS (Content-based Multimedia Retrieval System)**
  A modular and efficient client-server architecture for content-based retrieval of multimedia databases.

- **DAN (Distributed Agent Network)**
  A software agent-based distributed processing architecture for computer-telephony integration which provides ubiquitous service access. This is being developed in collaboration with VTT Electronics.

- **Beethoven**
  An architecture which offers a generic interface and gateway components for implementing open standards applications between an H.323 compliant IP telephony network and multimedia clients.

- **Princess**
  An architecture which produces scalable end-to-end services to mobile users by adapting the presentation according to the properties of the terminal type and the underlying network in question. This is being developed in collaboration with VTT Electronics.

Our research has brought forth a total of approximately 50 publications.

Exploitation of Results

The results of the algorithm level research and the functionality of the service platforms have been demonstrated with a number of successful pilots, for example the following:

- **A visual surveillance service built on top of the Princess architecture**
  The service monitors the live input of surveillance cameras, performing real-time motion detection. In case motion is observed, an alarm is stored into the alarm database and the clients who have subscribed to the service are notified by for example SMS or email.

- **IVCC (Internet Video Call Center) realized with the DAN architecture**
  The IVCC is based on the functionality provided by the DAN terminal and protocol abstraction, call control, data sharing, intelligent routing, video conferencing, and so on. The IVCC offers users with different terminals ubiquitous access to services provided either by IVCC agents or by human attendants.

Collaboration with partners

We value close collaboration with our various partners as an essential factor in the realization of research results into practical solutions. The most important partners of MediaTeam are the following funding bodies, companies, and research organizations:

The National Technology Agency; The Academy of Finland; Nokia; Acta; CCC; Sonera; OPOY/Finnet group; Sampo; Kesko; VTT Electronics; Mobile Forum Oulu; Language and Media Processing Laboratory, Univ. of Maryland, USA; Creative Media Lab, Univ. of Gävle, Sweden; National Institute of Standards and Technology, USA; and CEC Karlsruhe, SAP, Karlsruhe, Germany.

Future Goals

Media Team will continue to further develop its present status as a specialist and leading authority in multimedia research. We will pursue our goals of providing practical, easily adaptable solutions to the multimedia application industry and of performing eminent research in our fields of expertise. We will remain in active contact with the industry as a whole, as well as our existing partners, and strengthen and develop our international contact network which we regard as one of our essential assets. We will further look into our innovation process, to ensure an even more efficient manner of transforming research results into the practical applications of our customers.
Background and Mission

Machine vision is a continuously growing area of research dealing with processing and analyzing of image data. It plays a key role in the development of intelligent systems. The principal goal of research on intelligent systems is to develop machines that have the ability to perceive, reason, move, and learn from experiences.

The mission of the MVIS group is to carry out leading edge long term research on machine vision and intelligent systems technology, with an aim of bringing the results of the research close to applications.

The main areas of research are:
- Image analysis
- Image sequence analysis
- Intelligent systems

The research on image analysis concentrates on problems in texture analysis, color and face image analysis, document image analysis, and adaptive systems for color-texture inspection. The focus of image sequence analysis is in tracking and motion estimation, video indexing, camera calibration, and interventional MRI. The research on intelligent systems concentrates on context-aware mobile systems, intelligent service robots, neural network methodology for signal analysis, and analysis of biomedical ECG and EEG signals.

The research in the above areas provides the basis for original applied research in various areas, including industrial automation, mobile communications, multimedia, intelligent interfaces and medical engineering.

The group co-operates with many international and domestic partners. In applied research, the group has played central roles in European Esprit projects and several joint projects funded by the National Technology Agency (Tekes) and industry. As a result of our co-operation with the respected Fraunhofer IITB Institute (Karlsruhe, Germany), a one year postdoctoral visit to Fraunhofer was made by Dr. Hannu Kauppinen. The group is also active in the scientific community. For example, in 2000 the group organized an Infotech Workshop on Real-Time Image Sequence Analysis, Prof. Röning was elected a Fellow of the SPIE, and Prof. Pietikäinen was invited to the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *Pattern Recognition* journals.

The activities of the MVIS group are led by professors Matti Pietikäinen (Director), Juha Röning (Associate Director), Olli Silvén and Tapio Seppänen.

Scientific Progress

Image analysis

Image texture analysis has been a topic of intensive research for about three decades. During the past few years, the group has developed theoretically and computationally very simple nonparametric approaches to texture analysis based on Local Binary Patterns (LBP) and signed gray level differences. In 2000, LBP-type operators for multiresolution gray scale and rotation invariant texture classification were further investigated and research on the theoretical foundations of the rotation-invariant LBP was begun. Combined use of color and texture in classification was studied, with an application in the classification of parquet defects. The development of a Matlab toolbox for texture analysis using our operators was in progress.
Techniques for color-based face tracking in varying illumination conditions have been investigated. Our approach makes use of a chromaticity constraint called skin locus to select pixel candidates for adapting the face color distribution to the color shift caused by illumination change. The chromaticity constraint is a function in the chromaticity space which defines possible object colors in a certain color temperature range. The technique has been tested on several videos and has been shown to be successful in situations in which non-adaptive methods and adaptive methods based on spatial constraint can totally fail. New research activity concerning face recognition in varying environments was begun.

In our research on greenhouse plant vitality it was found that key indicators can be determined from the reflectance spectra. These include the water content, photosynthesis activity, and cues on stress. The measurement instrument used is an imaging spectrograph with 14-bit dynamic range that enables measuring phenomena that previously were detected only via fluorescence spectrometry. The current goal is to find spectral indicators to discriminate between directly illuminated leaves and those in a shadow, and between central and peripheral regions of leaves. This information is needed to increase the reliability of vitality measurements.

In wood surface inspection, a non-supervised training principle has been demonstrated to result in an essential improvement of defect detection and classification accuracy. At the lower level, the approach relies on our previous color-based inspection techniques, and the use of new texture methods is being investigated. The development of an easy to use software package for demonstrating the results is in progress. It is planned to be distributed to system integrators, end-users, and researchers in the field.

In document image analysis, an experimental system for automatic ground-truth generation was developed in order to evaluate the skew-tolerance of page segmentation methods. It is currently publicly available via the Internet. Research on extracting text information from document images with complex colored and textured backgrounds was begun. Simple texture-based methods using edge information were under development, providing a very promising performance in the preliminary experiments.

**Image sequence analysis**

A novel method for detecting and tracking humans based on particle filters has been developed. The method utilizes motion as a cue for segmenting objects in a scene into foreground or background. The camera can be either stationary or it can be moving. The basic idea is to estimate multiple motions from image sequences including also the camera motion. Unlike most of the tracking methods, the new approach does not require any explicit image features such edges or contours. Due to this property, the method could be suitable for many practical applications such as visual surveillance and robot navigation.

Examples of recent work in human tracking utilizing different modalities (left: color, middle: motion, right: shape).
In video coding, transform domain motion estimation techniques have been investigated. The main goal is to develop solutions that are computationally efficient and require low power consumption in hardware implementations. Several transform schemes have been tested using software implementations. The results indicate that there seems to be much potential in the new approach. Furthermore, a content based rate-control technique has been developed for video coding purposes. This technique is especially well suited for low bit-rate mobile applications with wireless communication channels.

The group has developed SOM (self-organizing map) based tools for video classification and indexing. A graphical user interface for training and visualizing the layered SOM structure is implemented with Java. The underlying video processing has been designed to extract color, motion and audio features from the video bitstream without completely decoding it. In this manner, the solution becomes computationally more efficient, which is important when dealing with large video databases.

The Matlab toolbox we have developed for geometric camera calibration was revised. A new and more powerful algorithm has been created which directly utilizes the geometry of the known control points. A variant of the toolbox has been implemented for geometric calibration of the line-scan cameras. Also a method for detecting and estimating the pose of the electrical components on a printed circuit board has been developed to be used in conjunction with the calibration toolbox.

**Intelligent systems**

The research on context-aware systems concentrates on software architecture, 3-D sensing technology, machine vision, context recognition and control methods. Distributed software architecture is being developed for mobile context-aware systems. The architecture offers well defined and reusable interfaces for different resources like sensors, actuators, computing devices and user interfaces. The location and implementation of a software component is transparent for the rest of the system.

The architecture is being applied both to systems controlling mobile service robots and to ubiquitous systems serving mobile users. In 2000, the work continued on basic robot resources like motion control, range sensors, dead-reckoning and vision. Work on a system serving a mobile user was begun; the basic architecture was improved and a software component offering an augmented reality user interface was developed.

A robust color tracking system was developed for a mobile robot, utilizing the expertise of the group in color vision. The system is capable of tracking objects of different colors in various lighting conditions. With this tracking system, a mobile system can focus its attention on the colored objects in the scene, and track a person based on skin color as well. For 3-D environment modeling, an experimental system is being developed which combines 3-D information provided by structured lighting with 2-D information provided by intensity (color) images. During this reporting period, a graphical user interface was developed. The next step is to mount the system in a mobile robot.

Our recent work on control methods has focused on evolutionary methods. Novel methods are being studied to construct mobile robot controllers using artificial neural networks. The methods use a dynamic topology of neurons for adapting to the environmental changes. Although a lot of work has to be done, the preliminary
results are promising. One of the main research challenges is the scalability of the methods to create an intelligent robot. To support this study, the development of a novel robot simulator has begun. To be able to utilize the group’s strong background in machine vision, a virtual color camera was implemented as one of the robot’s sensors.

Previous work on situated control architecture was continued. Our control system for playing simulated soccer acted as a qualifying competitor for the RoboCup’99. A student group participated in the simulation league of the fourth world cup of robotic soccer, RoboCup’00, held in Melbourne, Australia. In this competition the games of the simulation league were shown to the audience using the 3D visualization software developed by the student group.

In context recognition, a data set was collected from the actions of a user, as well as from the environment the user was in. The data set was preprocessed and different methods were tested for detecting the context of the user. Analysis was done by statistical signal processing techniques and neural networks. Neural networks were applied in other areas of signal analysis as well; in developing spatially and temporally adaptive learning techniques. Such techniques are needed as a static and deterministic system cannot cope with all the temporally changing situations in, for example, industrial environments or notably dynamic environments, such as the human physiological system.

New data mining methods are being developed for steel plant quality control. Data mining contains three closely related workflows: data preprocessing for data mining study, modeling of temperatures and dimensions after the finishing mill, and furnace control data study. The data preprocessing aims at finding variables and data items that make a significant contribution to defective coils. This study is carried out using statistical analysis, correlation study and the self-organizing maps. The modeling study strives to predict end product quality parameters with neural networks. The furnace control study surveys the dependencies between furnace control variables and develops new adaptive furnace control algorithms. The analysis is done using graph theory and neural networks.

Research into ECG signals has been conducted. The main goal in the project is to study heart rate variability of patients and healthy people in order to develop methods for automatic diagnosis and prognosis for clinical use. The medical part of the research is performed in the Oulu University Hospital while our group is responsible for signal analysis algorithms. The work is highly international, including universities from Europe and the USA. During 2000, an invited review paper on heart rate variability methods was published in an international scientific journal, and many journal papers applying our algorithms to clinical data were published.
New research on EEG signals has been launched in 1999. The goal of this study is to develop methods for automatic assessment of the depth of anaesthesia and physical reaction to pain in clinical operations. This goal is approached by performing adaptive segmentation and classification of evoked responses and phase synchronization of cortical cell populations. During 2000, five conference papers and two journal articles were published. The medical part of this research is conducted at the Oulu University Hospital while our group is responsible for algorithm development. The work involves other Finnish universities and industrial companies.

**Exploitation of Results**

The results of our research were applied to real-word problems in many projects, often in collaboration with industrial and other partners. Some examples of exploitation are described below.

Inspection methods developed by the group have been used in the wood grading system being developed for VTT Building Technology, and the non-supervised learning and user-interface approach developed by the group is used in the 3-D steel surface inspection systems made by Thermo Radiometrie.

The expertise of the group has been used in the development of an MPEG-4 codec solution by Hantro Products. The codec is intended to be integrated into wireless communications devices.

The European Esprit project called “Intraoperative real-time visualization and instrument tracking for MRI” (IRVIT) carried out in 1998-99 demonstrated the feasibility of an interactive MRI intervention system, which features real-time visual feedback. A follow-up project entitled “Advanced Minimally Invasive Therapy Using MRI” (AMIT) targeting clinical trials was begun during the reporting period.

The group’s expertise in robotics was applied in developing a mobile robot for domestic help. A teleoperated robot serves as the remote eyes of the elderly and those who take care of them. During the reporting period, the main task was to develop teleoperation capabilities for the robot. A system for controlling the robot remotely over the Internet was implemented. The first successful usability tests were performed. The work was carried out as a part of a project the goal of which is to create new product and service concepts for helping the elderly and disabled in their everyday lives. As a spin-off of this work, the development of a wheeled walking aid offering services based on modern information technology was started during 2000. The services envisaged for such a system include video calls, safety monitoring and controlling the environment.

**Future Goals**

In December 2000, the MVIS group was chosen by the Academy of Finland for the second round in the competition for national centre-of-excellence status and funding. For the final application submitted in February 2001, the group prepared its long-term research and operation plan covering the period 2002-2007. The group will continue to strengthen its long term research and researcher training. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects.
### Personnel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>11</td>
</tr>
<tr>
<td>graduate students</td>
<td>25</td>
</tr>
<tr>
<td>others</td>
<td>28</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>64</strong></td>
</tr>
<tr>
<td>person years</td>
<td>42</td>
</tr>
</tbody>
</table>

### External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>FIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>1 466 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>1 381 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>3 280 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>1 018 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>567 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>7 712 000</strong></td>
</tr>
</tbody>
</table>

### Doctoral Theses


### Selected Publications


Background and Mission

Machine vision is a continuously growing area of research dealing with processing and analyzing of image data. It plays a key role in the development of intelligent systems. The principal goal of research on intelligent systems is to develop machines that have the ability to perceive, reason, move, and learn from experiences.

The mission of the MVIS group is to carry out leading edge long term research on machine vision and intelligent systems technology, with an aim of bringing the results of the research close to applications.

The research on machine vision focuses on problems in texture analysis, color and face image analysis, document image analysis, tracking and motion estimation, 3D modeling and camera calibration, and visualization-based user interfacing. The research on intelligent systems concentrates on context-aware mobile systems, intelligent service robots, neural network methodology for signal analysis, and analysis of biomedical ECG and EEG signals.

The research in the above areas provides the basis for original applied research in various areas, including visual inspection, mobile communications, multimedia, intelligent interfaces and medical engineering.

The group co-operates with many international and domestic partners. In applied research, the group has played central roles in European Esprit projects and several joint projects funded by the National Technology Agency (Tekes) and industry. The group and its members are active in the scientific community. For example, in 2001 the group co-organized with the MediaTeam group an international Infotech Workshop on Information Retrieval (IR 2001), Prof. Matti Pietikäinen served on the editorial boards of two major journals, and several members of the group were on the committees of international conferences.

Dr. Janne Heikkilä won the prestigious young investigators prize of Tekniliiton edistämissäätiö (the Finnish Foundation for Technical Advancement) and senior investigator’s prize of the University of Oulu, Heikki Pylkkö won the award for the best Finnish Master’s thesis in pattern recognition in 2000 (Pattern Recognition Society of Finland) and Matti Niskanen the award for the best Master’s thesis in machine vision in 2000 (Vision Club of Finland). Our paper on unsupervised training in visual inspection by Matti Niskanen, Olli Silven and Hannu Kauppinen received the best paper award in the OCAV 2001 conference in France.

The activities of the MVIS group are led by professors Matti Pietikäinen (Director), Juha Röning (Associate Director), Olli Silvén and Tapio Seppänen.

Scientific Progress

Machine vision

Image texture analysis is an important fundamental problem in machine vision. During the past few years, the group has developed theoretically and computationally simple but very efficient nonparametric approaches to texture analysis based on Local Binary Patterns (LBP) and signed gray level differences. In 2001, LBP-type operators for multiresolution gray scale and rotation invariant texture classification were further investigated. A paper describing this method was accepted for publication in the prestigious IEEE Transactions on Pattern Analysis and Machine Intelligence journal. Joint use of color and texture in classification was studied, suggesting separate processing of color and pattern information. A real-time color and texture based method for defect recognition was also developed. A new framework for empirical evaluation of texture analysis algorithms is under development. For this purpose, a very large texture image database, Outex, was created containing over 300 different textures imaged with three different illuminants, nine rotations and six spatial scales (http://www.outex.oulu.fi).

Techniques for color-based skin detection and face tracking in varying illumination conditions were further investigated. Our approach makes use of a chromaticity constraint called skin locus to select pixel
candidates for adapting the face color distribution to the color shift caused by the illumination change. The skin color was not only studied on images; also spectral knowledge of it was utilized. The skin color signals, containing information both about reflectances of skin and about spectral power distributions of illumination, were also studied. A method combining skin locus based face modeling with the mean shift algorithm was developed for color-based face tracking. To make a reference for comparing different tracking methods, a new face video database was created with different cameras. It contains face videos with drastic color changes, color images with different camera calibrations and illuminations, ground truths for face localization, and spectral data.

A new approach to face detection in color images using a skin locus model and hierarchical filtering was developed. The method provides promising results in varying conditions, allowing the sizes and orientations of faces to differ in varying backgrounds and illumination. Approaches to face recognition were also investigated and some algorithms were implemented. When looking for issues to improve the face learning process, the properties of a new unsupervised learning approach, locally linear embedding, were investigated. The locally linear embedding algorithm gave also promising preliminary results in a visual inspection task.

In our research on grain size distribution determination, the purpose is to develop techniques for on-line grain size measurement in industrial processes. Methodology based on texture analysis and a non-segmenting approach is being developed. A new texture analysis method can efficiently discriminate different grain classes from each other.

In the research on plant vitality measurement, our concept of visualization-based user interfacing for inspection applications was adopted and further developed. With plants, the concept is used for investigating the effect of different features in spectral image segmentation. For observing the daily variations of plants, more test material was collected in a greenhouse environment.

In document image analysis, research on extracting text information from document images with complex colored and textured backgrounds was carried out. A simple texture-based method for text localization using edge information was proposed, as well as a method for finding horizontal and vertical text lines from complex binarized document images. New research concerning text extraction from scene images was begun.

In 2001, the research on image sequence analysis was focused on human tracking methods, transform domain video analysis, video coding, video surveillance,
Human tracking based on dominant motion extraction using particle filtering has been further developed. A real-time software module of this method was implemented for a PC platform as a part of a framework for a general purpose environment for visualizing and testing different tracking methods. Also, various regularization techniques have been explored for improving the spatial coherence of the human tracking results. The most prominent application area of this research is visual surveillance, but also other applications, including video coding, has been investigated.

In transform domain video analysis, an extensive amount of image transforms have been studied for several purposes. The main interests have been in block motion estimation and in object recognition. For motion estimation, a novel solution based on number theoretic transform has been developed. A highly optimized software module has been implemented for a general purpose processor that can be used in conjunction with a video encoder. A hardware implementation of this method is currently under development. Also, some other transforms have been found to be useful in motion estimation. For object recognition purposes, a completely new group of image transforms was proposed that can be invariant with respect to different linear transformations including affine transformation.

In video coding, a content based rate control technique has been developed which is able to utilize classification between foreground and background objects for adjusting the compression quality according to the corresponding priorities and the channel capacity. The new solution is compatible with most of the current video coding standards and it is also transparent for the decoder, and thus, only the encoder has to be modified.

The research carried out in video surveillance is closely related to video coding work. The main objective in this area is to develop solutions for wireless video surveillance systems. As a result, a framework for a stand-alone system has been produced, and a test system has been implemented with capabilities for preselecting, compressing, and wirelessly transmitting high resolution video, for example, from a moving vehicle to a remote control centre.

In video indexing, a Java based software package Video Data Mining System (VDMS) has been implemented for content based video retrieval. The system is capable of analyzing MPEG video clips based on certain feature data that are computed directly from the video stream. Different types of videos, such as music, sports and news, are then indexed using this information, and the user can make interactive visual searches of these videos with the aid of a new kind of a user interface, which represents the video database as two dimensional maps where similar videos are located in proximity to each other.

In video quality measurement, a novel solution for analyzing the video transmission quality has been developed. It can be used for measuring the effects of the bit errors in digital video, and to indicate the needs for improving the level of error correction. The solution used does not require any reference or test video to be transmitted but it can basically utilize the normal transmission mode without knowing the original source. The tests have shown that the quality measure obtained is comparable with the traditional signal to noise ratio measurement using known source data.

Camera calibration methodology has also progressed during this period. A new calibration procedure based on a coplanar target model and a linear algorithm has been proposed. A common problem of camera calibration in industrial applications is the need for large calibration targets that are difficult to make and maintain in a known three dimensional shape. Using this new method it is possible to utilize targets where the control points are located on a single plane, and still to achieve good precision of the camera parameters in a short period of time, whereas the traditional methods would have required much more computation.

Intelligent systems

The research on intelligent systems concentrates on context-aware mobile systems, intelligent service robots, neural network methodology for signal analysis, and analysis of biomedical ECG and EEG signals.
The research on context-aware systems focuses on software architecture, 3-D sensing technology, machine vision, context recognition and control methods. Distributed software architecture is being developed for mobile context-aware systems. The architecture offers well defined and reusable interfaces for different resources like sensors, actuators, computing devices and user interfaces. The location and implementation of a software component is transparent for the rest of the system.

The architecture is being applied both to systems controlling mobile service robots and to ubiquitous systems serving mobile users. In 2001, a general agent-based architecture, Genie of the Net, for managing services on behalf of the user was developed. A first prototype, the Family Calendar, was completed and demonstrated. This work was in co-operation with VTT Electronics. The work on mobile robots continued on basic robot resources like motion control and vision. This was realized in a new, Finnish Academy funded project “Robots Serving Humans”. The goal of this project is to develop methods and components for the next generation intelligent service robots. In this project, we emphasize the role of robust senses (especially color vision) and learning (self-organization). The main application, and thus the test-bed for the developed methods, is intelligent telepresence, where the robot is a semi autonomous agent providing a convenient way to monitor and access remote environments for a human user.

Evolutionary methods to evolve neural controllers for a mobile robot were studied. An approach called Evolutionary Neuron Migration was used to evolve neural control structures for the mobile robot. The neural structures were able to solve real problems in perception and control. The feasibility of the approach was demonstrated by evolving robust navigation behavior for a real robot.

A robust color tracking system has been developed for a mobile robot too, utilizing the expertise of the group in color vision. The system is capable of tracking objects of different colors in various lighting conditions. With this tracking system, a mobile system can focus its attention on the colored objects on the scene, and track a person based on skin color as well. The first successful experiments were performed with the color tracking system to control a soccer robot. The robot was able to find a moving ball and approach it. An article about the color tracking system was published in the IEEE International Conference on Robotics and Automation.

Previous work on situated control architecture was continued. Our control system for playing simulated soccer acted as a qualifying competitor for the RoboCup’99. A student group participated in the simulation league of the fourth world cup of robotic soccer, RoboCup’01, held in the USA. In this competition, the games of the simulation league were shown to the audience using the 3D visualization software developed by the student group.

The distributed software architecture was further developed. A common interface was developed for all robotics resources. Furthermore, methods to dynamically construct state machines for robot controllers were studied.

New data mining methods were developed for steel plant quality control. Data mining contains three closely related workflows: data preprocessing, steel strip quality prediction, and furnace control data study. The data preprocessing aims at finding variables and data items that make a significant contribution to defective coils. The methods for this purpose were statistical analysis, visualization of parallel coordinates, k-means clustering and self-organizing maps. The steel strip quality study utilized neural networks for predicting the temperatures and dimensions of steel strips after the finishing mill. The modeling was done with the information available before the finishing mill phase. In the furnace control data study, an adaptive neural network model was built for predicting the post roughing mill temperature of the pre-rolled steel strip. The furnace control software was written based on the results of the study. The development of an intelligent furnace control system continues.

Research into ECG signals has been conducted. The main goal in the project is to study heart rate variability of patients and healthy people in order to develop methods for automatic diagnosis and prognosis for clinical use. The medical part of the research is performed in the Oulu University Hospital, while our group is responsible for signal analysis algorithms.
The work is highly international, including universities from Europe and the USA. Due to the success in the scientific research, the ECG team was granted the 3-year status of Quality Research Unit of the University of Oulu. This enabled the opening up of a new research area, the analysis of T-wave dynamics and morphology. During 2001, three journal papers on heart rate variability methods were published, and four journal papers were submitted for evaluation.

Research on EEG signals was launched in 1999. The goal of this study is to develop methods for automatic assessment of the depth of anesthesia and physical reaction to pain in clinical operations. The work is performed in cooperation with the Oulu University Hospital and our EEG team is responsible for algorithm development. During the year 2001, research was directed towards inclusion of other types of biosignals to the analysis system, too, and performing sensor fusion for estimating better the above mentioned parameters. During 2001, one journal article was published and two were accepted for publication.

Inspection methods developed by the group have been used in the wood grading system being developed for VTT Building Technology, and the non-supervised learning and user-interface approach developed by the group is used in the 3-D steel surface inspection systems made by Thermo Radiometrie.

The European Esprit project called “Intraoperative real-time visualization and instrument tracking for MRI” (IRVIT) carried out in 1998-99 demonstrated the feasibility of an interactive MRI intervention system, which features real-time visual feedback. A follow-up project “Advanced Minimally Invasive Therapy Using MRI” (AMIT) targeting clinical trials was in progress during the reporting period. A new proposal on exploiting part of the results in the form of open software was in preparation.

The group’s expertise in robotics was applied in developing a mobile robot for domestic help. A teleoperated robot serves as the remote eyes of the elderly and those who take care of them. During the reporting period, the main task was to develop teleoperation capabilities for the robot. A voice controlled service robot was successfully demonstrated. The purpose of the robot is to assist elderly people in their homes and provide a communication link to the health care personnel. A design project was launched with the University of Lapland to further develop the appearance of the robot, and make it suitable for various applications and research studies regarding human-robot interaction.

The development of a wheeled walking aid offering services based on modern information technology has continued during 2001. The project is in cooperation with the University of Lapland. The services envisaged for such a system include video calls, safety monitoring and controlling the environment.

**Future Goals**

We will continue to strengthen our long term research and researcher training. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects. In order to further sharpen its focus, the MVIS group has been recently divided into two separate but closely cooperating groups: Machine Vision (http://www.ee.oulu.fi/mvg)
and Intelligent Systems (http://www.ee.oulu.fi/research/isg). These groups together with our earlier “spin-off” group MediaTeam Oulu (http://www.mediateam.oulu.fi) form the Machine Vision and Media Processing Unit (MVMP) (http://www.ee.oulu.fi/mvmp). The MVMP unit coordinates research and researcher training activities between its groups and has some joint facilities and support staff.

**Personnel**

<table>
<thead>
<tr>
<th>Personnel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>12</td>
</tr>
<tr>
<td>graduate students</td>
<td>22</td>
</tr>
<tr>
<td>others</td>
<td>32</td>
</tr>
<tr>
<td>total</td>
<td>66</td>
</tr>
<tr>
<td>person years</td>
<td>48</td>
</tr>
</tbody>
</table>

**External Funding**

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>410 700</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>302 400</td>
</tr>
<tr>
<td>Tekes</td>
<td>780 700</td>
</tr>
<tr>
<td>domestic private</td>
<td>355 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>180 100</td>
</tr>
<tr>
<td>total</td>
<td>2 029 000</td>
</tr>
</tbody>
</table>

**Doctoral Theses**


**Selected Publications**


Background and Mission

Machine vision is a continuously growing area of research dealing with processing and analyzing of image data. It plays a key role in the development of intelligent systems. The principal goal of research on intelligent systems is to develop machines that have the ability to perceive, reason, move, and learn from experiences.

The mission of the MVIS group is to carry out leading edge, long term research on machine vision and intelligent systems technology, with an aim of bringing the results of the research close to applications.

The research on machine vision focuses on methods and systems for analyzing images and image sequences. The aim is to develop new approaches, which provide the leading performance, are robust with respect to changes in environment and are computationally efficient. The research on intelligent systems concentrates on mobile and context-aware systems, data mining methods and secure programming. Our aim is to increase systematically our methodological understanding and to develop components for an intelligent environment that gives versatile services for its inhabitants.

The research in the above areas provides the basis for original applied research in various areas, including visual inspection, mobile communications, multimedia, intelligent interfaces and medical engineering.

The group co-operates with many international and domestic partners. In applied research, the group has played central roles in European Esprit projects and several joint projects funded by the National Technology Agency (Tekes) and industry. The group and its members are active in the scientific community. For example, in 2002 the group co-organized with the MediaTeam group an international conference, MUM 2002 (1st International Conference on Mobile and Ubiquitous Multimedia, 11-13 December); Prof. Matti Pietikäinen served on the editorial boards of two major journals; and several members of the group were on the committees of international conferences. Olga Kouropteva won the award for the best Finnish Master’s thesis in pattern recognition in 2001 (Pattern Recognition Society of Finland).

The activities of the MVIS group are led by professors Matti Pietikäinen (Director), Juha Röning (Associate Director), Olli Silvén, Janne Heikkilä, Jukka Riekki and Tapio Seppänen.

Scientific Progress

Machine vision

Image texture analysis is an important fundamental problem in machine vision. During the past few years, the group has developed theoretically and computationally simple but very efficient nonparametric approaches to texture analysis based on Local Binary Patterns (LBP) and signed gray level differences. In 2002, the joint use of color and texture in classification was studied, suggesting separate processing of color and pattern information. Outex, the new large texture image database created for empirical evaluation of texture analysis algorithms was further developed (http://www.outex.oulu.fi). The performances of the MPEG-7 texture descriptors and the LBP method were empirically evaluated in the retrieval of 319 Outex textures. LBP clearly outperformed the MPEG-7 descriptors both with respect to the retrieval accuracy and the computation time. Multiscale versions of the LBP operator were studied. A novel method, based on cellular automata, was proposed for this purpose. Research on view-based recognition of 3D textured surfaces was begun. Our method provided the leading performance for CURET textures and very promising results for outdoor scene images. A method for real-time surface inspection according to texture was developed, as well as a methodology for optimizing color and texture features for real-time inspection applications.

Techniques for color-based skin detection and face tracking under drastic illumination changes were further investigated. Our approach is based using a chromaticity constraint called a skin locus to select pixel...
candidates. This method has been applied to both image sequences and still images. In the first case, the skin locus is used for adapting the face color distribution. The skin locus performed very well in a comparison with other state-of-the-art methods that was carried out for both sequential and still images. Research concerning the distributions of all 8-bit RGB values in different color spaces was performed. These color spaces were also tested for the use of the dichromatic model and mixtures of light sources and skin spectra.

Research on the recently proposed unsupervised learning and dimensionality reduction method, locally linear embedding (LLE), was continued. A nonparametric supervised version of LLE was developed. In cooperation with the Pattern Recognition Group of the University of Delft, the methods of supervised LLE independently developed by both groups were unified in a common framework and applied to a number of benchmark data sets. The sensitivity of the original unsupervised LLE to outliers was also studied and a robust extension to tackle this problem was proposed.

The visualization-based approach for surface inspection was further investigated. The method combines an intuitive user interface with an unsupervised classifier. The idea is to map the unlabeled training data into two dimensions with some dimensionality reduction method, like the self organizing map (SOM). Then the training data is visualized on the 2D map where the user determines the boundaries between classes. The approach does not require labeling of separate samples, which is often an inconsistent and error prone task reducing the accuracy achieved, but the decision of the class boundaries is based on observing the data set as a whole. With this approach, the retraining for changing material or imaging conditions involves simply moving the decision boundaries on the 2D map. The approach is also computationally attractive for real-time applications.

Research on visualization-based paper characterization according to texture was in progress. A new methodology based on combining LBP-based texture feature extraction with unsupervised training and classification using SOM was developed. The preliminary results were excellent, reducing the classification error to under a fortieth part of that obtained with previous methods. In addition to the excellent classification accuracy, the method also offers a self-intuitive user interface and a synthetic view to the inspected data.

Our research on grain size distribution determination is based on using a non-segmenting approach and unsupervised classification methods. Image grain size information is extracted with efficient texture features like LBP and spectral information. Two and three fraction mixtures have been tested to evaluate the performance of the features. The purpose is to develop techniques for on-line grain size measurement in industrial processes.

In the research on plant vitality measurement, the SOM classification method, with several spectral features as input, was used to discriminate between plants with different vitality characteristics. Other dimensionality reduction methods, like LLE, are planned to be considered in future research.

The video processing research was focused on motion estimation and video stabilization. The main objective in motion estimation was to develop a hardware module for performing computationally the most intensive part of video coding more efficiently without loss of accuracy. As a result, a low-power solution for real-time block motion estimation for CIF-sized frames was designed that utilizes a novel transform based approach. Another result in motion estimation research was an improved method for eliminating poor candidate vectors in spatial domain full search block motion estimation with the sum of squared differences (SSD) error criterion.

In video stabilization, a new framework was developed. This framework is mainly targeted for mobile devices with video encoding capabilities. The feasibility of this approach has been successfully demon-
strated with a software implementation and a large set of test material. The results have shown that the video quality has improved significantly while the video bit-rate has dropped substantially due to the attenuation of camera vibrations. The main advantages of the approach developed are the optimal separation of the intended and unintended motions, and the small overhead needed when stabilization is used together with a video encoder.

In transform domain pattern analysis, the properties of the Multi-scale autoconvolution (MSA) transform developed in our group were studied. In practical applications, the transform requires some preprocessing of the images and selection of the appropriate set of transform coefficients. This research is still in progress. Other possible image transforms have also been investigated to achieve necessary invariances against geometric transformations, and to solve the alignment problem between pairs of images under various geometric distortions.

Pattern recognition using multi-scale autoconvolution.

New research on machine vision for sensing and understanding human actions was begun. Our project proposal on this topic was accepted for the Proactive Computing program of the Academy of Finland. The general objective is to investigate the capabilities of machine vision in proactive computing and to develop solutions needed for building emerging applications.

An embedded vision module is planned to be developed using vision software architecture based on event analysis. It will be provided with machine vision skills for sensing and identifying humans and their actions in varying conditions and with capabilities to wirelessly communicate with other units. The applicability of the approach will be shown in an intelligent room equipped with multiple cooperating vision units, demonstrating complex capabilities built on simple modules. A wireless local area network connected to the Internet is used for communication and database access.

**Intelligent systems**

The research on intelligent systems concentrates on mobile and context-aware systems, data mining methods and secure programming. These are applied in context-aware mobile systems, intelligent service robots and analysis of biomedical ECG and EEG signals.

The research on context-aware systems focuses on software architecture, 3D sensing technology, machine vision, context recognition and control methods. Distributed software architecture is being developed for mobile context-aware systems. The architecture offers well defined and reusable interfaces for different resources like sensors, actuators, computing devices and user interfaces. The location and implementation of a software component is transparent for the rest of the system. The architecture was further developed. A common interface was developed for all robotics resources. Furthermore, methods to dynamically construct state machines for robot controllers were studied.

The architecture is being applied both to systems controlling mobile service robots and to ubiquitous systems serving mobile users. In 2002, development of a general agent-based architecture, Genie of the Net, for managing services on behalf of the user has progressed. This work has been done in co-operation with VTT Electronics. The work on mobile robots continued with basic robot resources like motion control and vision. This has been realized in the Finnish Academy funded project “Robots Serving Humans”. The goal of this project is to develop methods and components for the next generation intelligent service robots. In this project, we emphasize the role of robust senses (especially color vision) and learning (self-organization). The main application, and thus the testbed for the developed methods, is intelligent telepresence, where the robot is a semi autonomous agent providing a convenient way to monitor and access remote environments for a human user.
As part of the future environment, we envisage home robots that can be used for different purposes, such as telepresence. In our research laboratory, we are developing a robot whose main purpose is to serve as a telepresence body, but which can also perform home aid functions. The goal is to develop new methods and devices that will improve the current telepresence equipment by providing a self-guiding, intuitive, and very easy-to-use user interface and by creating a realistic feeling of being present in the remote location. During the reporting period, work on speech recognition over a videoconference connection and robot sensor visualization via the video stream of the videoconference has been done. Our techniques allow the robot to be remotely controlled with generally available videoconference software applications such as Microsoft Netmeeting.

Evolutionary methods to evolve neural controllers for a mobile robot were studied. An approach called the Stochastic Evolutionary Neuron Migration Process (SENMP) was used to evolve neural control structures for mobile robots and to gain new insights into adaptation in neural networks. The neural structures were able to solve real problems in perception and control. The feasibility of the approach was demonstrated by evolving robust navigation behavior for a simulated and a real mobile robot.

A graph theoretical approach was studied for creating dynamic robot formations. The approach uses a leader-follower scheme, allowing multiple robots to move in formations and change the formations when necessary due to environmental conditions.

A new control architecture called Bender has been developed for mobile robots. It is developed as a general architecture for robots in our laboratory. Bender is a real-time, dynamic, and modular control architecture. It is the most recent member in the family of control architectures developed in our laboratory, using features from SAMBA and COCOA and building on top of the distributed architecture discussed above. Bender was primarily designed to be a control architecture allowing easy implementation of new methods and algorithms. It also aims to be a general framework for different mobile robots, as it should work in robots of different kinds and sizes. Dynamic modules provide a flexible way for creating different control systems from fully reactive to deliberative ones. The modules are usually created when the system starts, but they can also be created and connected to other modules dynamically during execution for a special purpose. The modules receive and send information to other modules in a unified format, called markers. Markers are XML formatted structures that refer to physical objects or abstract things, and they are also used for creating an environment model core.

Bender is a part of the distributed software architecture providing a way to distribute modules into a network, as some of the modules can be started in other machines and accessed via CORBA interfaces implemented for all robotics resources, ranging from color cameras to motion control systems. Furthermore, a dynamic state machine architecture was developed that allows the user to create and modify online a complete state machine, utilizing all distributed robotics resources by simply writing and executing an XML formatted description of the state machine.

In the context-aware research, the group is also participating in the CAPNET research program. The purpose of CAPNET (Context-aware Pervasive Networking) research program is to create a foundation for new information and communications technologies for business in the field. Three research groups from the University of Oulu participate in the research program: MediaTeam from the Information Processing Laboratory, Intelligent Systems from the Computer Engineering Laboratory, and Human Computer Interaction from the Department of Information Processing Science.

The focus of the CAPNET program is on context-aware mobile technologies for ubiquitous computing. These are technologies, which allow communication anytime and anywhere, with any kind of terminal device, automatically taking into account the characteristics of the network and the terminal. The Intelligent Systems Group participates in the CAPNET program by studying context-awareness and developing software architectures. Context-awareness research concentrates on identifying the user’s routines utilizing location information acquired through WLAN, and a pressure sensitive floor on the premises of the Tietotalo Robotics Laboratory.
New data mining methods were developed for steel plant quality control and spot welding processes. The steel plant quality control study has continued. The two-year VALTA project in co-operation with Rautaruukki Steel in Raahe contains three application areas: furnace control study, predicting the scaling of the steel strips and breaking strength modeling for steel plates. The research project is partly a continuation of the earlier HiTech Steel project.

The furnace control study has progressed to intelligent furnace control system development. During this period, the usability of prior information was studied and the development of the steel slab temperature predicting system was continued. The impact of temperatures of the ceiling and heating time to the final slab temperature was modeled. The earlier neural network model predicted the temperature of the steel slab after the roughing mill, but during this research period the model has been improved to also predict the earlier temperatures.

The predicting of the steel strip scale defects started in June. In the preprocessing state, the relationships of the variables were visualized with scatter plots and self-organizing maps. The association rules were also used. The variables causing the scaling seem to interact widely. The research continues. A predicting model for the breaking strength of steel plates was developed. Linear, log-linear and neural network models were used. The planning model for the breaking strength of steel plates has been improved and the first Excel based table models are in the testing phase at Rautaruukki Steel in Raahe. The research will continue analyzing the deviation of the breaking strength, which will be estimated with heteroscedastic regression models.

In the research on applying data mining methods to steel industry measurements, we have also co-operated with Arizona State University where one of our researchers spent a year as a visiting researcher.

SIoux (Intelligent system for dynamic online quality control of spot welding processes for cross-sectoral applications) is a two-year EU-project developing a real time quality control system of resistance spot welding joints. The aim of the system is to replace time and cost intensive destructive tests by using soft computing methods for quality estimation. The benefit is that the approach does not require any additional sensors, like for example ultrasound based methods, as it uses on-line process data combined with knowledge gathered from previous experiments.
We are building a dynamic database containing measurements from welding tests and developing data mining methods for analyzing the measurement data to achieve better welding quality. During the last period the database has been partly set up and methods for pre-processing and analyzing the measurements have been developed.

The system is an application example of one of our research topics - a smart archive. A smart archive is a system combining expertise in databases and data mining methods. The goal is to implement and bring this know-how to processes that produce large amounts of data and demand adaptive, continuous and largely autonomous supervision.

Research into ECG signals has been conducted. The main goal in the project is to study heart rate variability of patients and healthy people in order to develop methods for automatic diagnosis and prognosis for clinical use. The medical part of the research is performed in the Oulu University Hospital, while our group is responsible for signal analysis algorithms. The work is highly international, including universities from Europe and the USA. One graduate thesis (M.Sc.) was finished in 2002 on T-wave morphology research. Computer software for analyzing T-wave dynamics was written and several patient data analyses were conducted. The software is presently in extensive use at the University Hospital of Oulu. One conference paper was submitted on this result. Heart rate variability was investigated and as a result, one journal paper on heart rate variability methods was published, and two journal papers were revision-submitted for evaluation.

Research on the methods for automatic assessment of the depth of anesthesia and physical reaction to pain was continued. A new multi-sensor approach to this is being developed. New patient material was measured and analysis software was written. The work is performed in cooperation with the Oulu University Hospital and the Lappeenranta University Hospital while our team is responsible for algorithm and software development. During 2002, two journal articles were published.

A new research topic of muscle symphatetic nervous activity (MSNA) signal processing was established. The aim is to study the basic relations between heart rate, blood pressure, breathing and MSNA signals. One M.Sc. thesis was completed on the topic in which computer software was developed for this purpose. Ph.D. work has already been started. Two conference papers were submitted for evaluation in 2002.

Inside the Intelligent Systems Group, the Oulu University Secure Programming Group (OUSPG) has kept its focus on implementation level security issues and software security testing. Software implementation may introduce potential for unanticipated and undesired program behaviour, e.g. an intruder can exploit the vulnerability to compromise the computer system. The group has researched different approaches to testing implementations of protocols using black-box (i.e. functional) testing methods (PROTOS). The recent discoveries of the group have raised awareness of security issues globally, e.g. test results in SNMP (Simple Network Management Protocol, http://www.cert.org/advisories/CA-2002-03.html). As a result of PROTOS research, a spin-off company, Codenomicon Ltd., was founded in 2001 to develop commercial software testing products (http://www.codenomicon.com).

Interdisciplinary research was carried out on ethical and communication issues related to software vulnerabilities. A vulnerability communication survey attracted over 150 international participants, with over 50 entries from software vendors and prior to the late public outbreak of debate on ethics of software security a journal article on the very topic was prepared.

### Exploitation of Results

The results of our research were applied to real-world problems in many projects, often in collaboration with industrial and other partners. Some examples of exploitation are described below.

In the wood inspection area, several board edgers using the visual training approach have been installed at sawmills by an industrial partner. These have become fully operational within hours of installation unlike earlier systems which required at least a several weeks training period.

The European Esprit project called “Intraoperative real-time visualization and instrument tracking for MRI” (IRVIT) carried out in 1998-99 demonstrated the feasibility of an interactive MRI intervention system which features real-time visual feedback. A follow-up project “Advanced Minimally Invasive Therapy Using MRI” (AMIT) targeting clinical trials was in progress during the reporting period. A start-up company has been founded by researchers, building on the foundation laid by the project.

The research in video sequence processing and analysis has contributed to recent video codec products that have been integrated into mobile communications devices. Although the key role of the researchers was mostly in enabling rapid product development, the result is a convincing demonstration of the benefits of long term research.
The group’s expertise in robotics was applied in developing a mobile robot for domestic help. A teleoperated robot serves as the remote eyes of the elderly and those who take care of them. During the reporting period, the main task was to develop teleoperation capabilities for the robot. A voice controlled service robot was successfully demonstrated. The purpose of the robot is to assist elderly people in their homes and provide a communication link to the health care personnel. A design project was launched with the University of Lapland to further develop the appearance of the robot, and make it suitable for various applications and research studies regarding human-robot interaction.

The development of a wheeled walking aid offering services based on modern information technology has continued during 2002. The project is in cooperation with the University of Lapland. The services envisaged for such a system include video calls, safety monitoring and controlling the environment.

During 2002 significant practical penetration for developed theories and procedures were achieved in the software industry and security community. First, home grown testing methodology was applied in a SNMP test-suite, where the OUSPG, CERT/CC and tens of vendors worked together on a model developed in OUSPG to address serious and widespread problems. Second, in cooperation with the MediaTeam, a test-suite was developed for the Session Initiation Protocol (SIP). This was introduced in a novel way and also well received in a global SIP Interoperability Meeting of protocol pioneers and vendors (October 2002).

### Future Goals

We will continue to strengthen our long term research and researcher training. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects. In order to further sharpen its focus, the MVIS group was recently divided into two separate but closely cooperating groups: Machine Vision (http://www.ee.oulu.fi/mvg) and Intelligent Systems (http://www.ee.oulu.fi/research/isg). Both of these were accepted to Infotech Oulu for the period 2003-2005.

### Personnel

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>12</td>
</tr>
<tr>
<td>graduate students</td>
<td>29</td>
</tr>
<tr>
<td>others</td>
<td>40</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>81</strong></td>
</tr>
<tr>
<td>person years</td>
<td>56</td>
</tr>
</tbody>
</table>

### External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>355 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>340 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>408 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>256 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>203 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>1 562 000</strong></td>
</tr>
</tbody>
</table>

### Doctoral Theses


Selected Publications

**Machine vision**


**Intelligent systems**


CERT® Advisory CA-2002-03 Multiple Vulnerabilities in Many Implementations of the Simple Network Management Protocol (SNMP) http://www.cert.org/advisories/CA-2002-03.html


Background and Mission

Machine vision is a continuously growing area of research dealing with processing and analyzing of image data. It plays a key role in the development of intelligent systems. The mission of the MV group is to carry out leading edge, long term research on machine vision technology, with an aim of bringing the results of the research close to applications.

The research on machine vision focuses on methods and systems for analyzing images and image sequences. The aim is to develop new approaches, which provide the leading performance, are robust with respect to changes in the environment and are computationally efficient. The research in the above areas provides the basis for original applied research in various areas, including visual inspection, mobile communications, multimedia, intelligent interfaces and medical engineering.

The group co-operates with many international and domestic partners. In applied research, the group has played central roles in European Esprit projects and several joint projects funded by the National Technology Agency (Tekes) and industry. The group and its members are active in the scientific community. For example, in 2003 Prof. Matti Pietikäinen served on the editorial boards of two major journals; and members of the group were on the committees of international conferences. Markus Turtinen won the Vision Prize 2003 awarded by the Vision Club of Finland for his Master’s thesis on texture-based paper inspection. The activities of the MV group are led by professors Matti Pietikäinen (Director), Olli Silvén (Associate Director) and Janne Heikkilä.

Scientific Progress

Texture, Color, and Face Analysis

Image texture analysis is an important fundamental problem in machine vision. During the past few years, the group has developed theoretically and computationally simple, but very efficient nonparametric methodology for texture analysis based on Local Binary Patterns (LBP). The approach has evolved to represent a significant breakthrough, outperforming earlier methods in many applications. It is already widely used all over the world. In 2003, a multiscale version of the operator, based on cellular automata was further developed. A visualization-based approach using self-organization of LBP feature distributions was proposed for training a classifier for textured scene images. A publicly available C++ library for image analysis has been developed by Dr. Topi Mäenpää, including implementations of our texture methods and many other useful image analysis and pattern recognition operations. Outex, the large texture image database created for empirical evaluation of texture analysis algorithms was further developed to contain outdoor scene images.

Scene image classification with local binary patterns. a) The original image. b) Ground-truth regions. c) Classified pixels within ground-truth regions. d) Segmented image.

The LBP methodology was adopted to the task of face recognition. In this approach, the face image is divided into several regions from which the LBP features are extracted and concatenated into an enhanced feature vector to be used as a face descriptor. In extensive experiments using FERET test images and protocol, considering variations in facial expressions, lighting and aging of the subjects, our methodology outperformed the state-of-the-art methods. A paper on this topic was recently accepted for the 8th European Conference on Computer Vision (ECCV 2004).
A compact LBP-based feature space was also developed for face detection and for the recognition of low-resolution face images. Also in these problems, our approach performed very favorably compared to the state-of-the-art. A paper on this topic will appear in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004).

The effects of face sequence length and image quality on the performance of the video-based face recognition systems using a spatio-temporal instead of a still image based representation were studied. An unsupervised approach for selecting representative face samples (models) from raw videos was also proposed and utilized for appearance-based face recognition.

Techniques for color-based skin detection under drastic illumination changes were further investigated. Our approach is based using a chromaticity constraint called a skin locus to select pixel candidates. During the reporting period, we continued our comparative study on skin detection methods and found that the skin locus method is more tolerable to illumination changes than other state-of-the-art methods. We also began to study the use of color information with scene images. Dr. Birgitta Martinkauppi visited one of the leading color experts, Prof. Finlayson and his group at the University of East Anglia, UK. An invited chapter on facial skin color modeling was also written for the “Handbook of Face Recognition” edited by S.Z. Li and A.K. Jain, demonstrating that our skin color research is highly respected internationally.

Research on the recently proposed unsupervised learning and dimensionality reduction method, locally linear embedding (LLE), was continued. This method was included in a new general framework for visualizing high-dimensional data. The framework provides a deeper insight into visualization. It consist of four components: metric learning, intrinsic dimensionality estimation, feature extraction, and feature selection. The framework was tested on several real-world datasets.

Research on visualization-based paper characterization according to texture was continued. Earlier, a new methodology based on combining LBP-based texture feature extraction with unsupervised training and classification using the self-organizing map (SOM) was developed. During the reporting period, methods for visualizing high-dimensional feature data were studied. The Isomap method gave very promising results providing useful information about the data in the training phase. Better comprehension of the properties of the paper and their effects on the final product were utilized in the paper formation analysis.

Our research on grain size distribution determination is based on using a non-segmenting approach and unsupervised classification methods. Image grain size information is extracted with efficient texture features like LBP and spectral information. The purpose is to develop techniques for on-line grain size measurement in industrial processes. In this context, the development of a solution for measuring the grain size distribution of stone material used in road asphalt production is beginning. The initial target is a practical system for the laboratory of an asphalt station.

The research on plant vitality measurement using imaging spectrography has employed SOM as the clustering and classification method, also other dimensionality reduction methods such as LLE and Isomap have been considered. A key finding is the
patchiness of leaves, indicating that photosynthesis operates in an on/off manner. This can be utilized in analyzer solutions.

**Video Analysis**

In the video processing research, our main objective has been to develop methods and solutions for improving video coding efficiency, as well as the image quality in wireless video transmission. The hardware solution for block-motion estimation up to ¼ pixel precision invented in our group gives us a basis for acquiring optimal motion vectors in the mean squared sense in real-time. The rate-distortion optimization method developed provides computationally an effective way of selecting those motion vectors that result in both low bit rate and high quality. Also, the motion vector field obtained becomes more uniform, revealing the underlying true motion better than before, which makes it possible to analyze the video contents more reliably directly from the motion vectors.

Another way of improving the video quality is to attenuate the unwanted vibrations from the video stream captured using a handheld camera. Our group has developed a new stabilization method for performing this operation in conjunction with the video encoding task. A new algorithm utilizing adaptive filtering and advanced motion analysis provides both robustness and fast performance for the method. The experiments conducted are very promising and they indicate that the method could be implemented into a mobile terminal in the near future.

Our research on visual tracking has resulted in a real-time head tracker that is capable of estimating the pose of the head of a human being in three dimensions from facial features. Currently, the features that are used in tracking are the eyes and the mouth. We have also investigated using this tracker as a mechanism for controlling user interfaces such as computer desktops.

**Geometric Invariants and Descriptors**

In the research on geometric invariants and descriptors, we have continued our investigations on the properties of the Multi-scale autoconvolution (MSA), which is an affine invariant transform developed in our group. As a result, we have been able to limit the transform space to a finite domain, which contains all the independent transform coefficients. We have also found that certain coefficients can provide information about the convexity of the objects. This gives us a basis for building simple classifiers to determine if the object in the image is convex and what is the degree of its convexity.

Another objective in our tracking research has been to develop an automated distance education assistant that is capable of selecting video sources from different cameras in an intelligent way.

**Proactive Machine Vision**

Research on machine vision for sensing and understanding human actions was in progress as a part of the Proactive Computing program of the Academy of Finland. A real-time head tracker based on facial features and continuous estimation of head pose was developed as described above. This research is being continued in a project funded by Tekes. Our aim is to develop a user interface for mobile devices based on head tracking. A distributed Corba-based architecture...
for proactive machine vision was studied and demonstrated in a system composed of three autonomous vision modules. A method for person identification based on color histograms of clothes was also developed. In addition, research on a novel texture-based approach to subtracting the background of moving objects was begun. Preliminary results with an LBP-based method were very promising compared to the state-of-the-art.

The user interface developed for proactive vision systems visualizes the data as a two-dimensional projection of a multidimensional feature space on the computer screen in such a manner that its topology is preserved. Self-organizing maps are used for this purpose in our implementation. Now, the user is able to locate the data clusters that represent meaningful or interesting events, and can assign a desired response to them when detected. The figure shows an experimental user interface where similar scenes are clustered to nodes that are close to each other.

**Exploitation of Results**

The results of our research were applied to real-world problems in many projects, often in collaboration with industrial and other partners. Some examples of exploitation are described below.

In the wood inspection area, several board edgers using the visual training approach have been installed at sawmills by our industrial partner INX Systems. These have become fully operational within hours of installation unlike earlier systems which required at least a several weeks training period. Recently, INX Systems obtained the ECVision (Excellence on Cognitive Vision Systems) Prize for its OptiGrader machine vision system which applies our research results.

Our method for on-line paper characterization has also been on trial use in several paper mills in Finland and abroad. Our industrial partners have expressed interest in utilizing visualization-based data analysis methods, like Isomap, in the training and further analysis of the paper texture data.

The research in video sequence processing and analysis has contributed to recent video codec products that have been integrated into mobile communications devices. Although the key role of the researchers was mostly in enabling rapid product development, the result is a convincing demonstration of the benefits of long term research.

**Future Goals**

We will continue to strengthen our long term research and researcher training. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects.

**Personnel**

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>10</td>
</tr>
<tr>
<td>graduate students</td>
<td>12</td>
</tr>
<tr>
<td>others</td>
<td>11</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>33</strong></td>
</tr>
<tr>
<td>person years</td>
<td>24</td>
</tr>
</tbody>
</table>
External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>154 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>179 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>183 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>192 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>1 500</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>709 500</strong></td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


Background and Mission

The Machine Vision Group of the University of Oulu was established in 1981. It has achieved a highly respected position in its field, and its results have been widely exploited in industry. The objective of the group is to carry out leading edge long term research on machine vision technology.

The wide-spread use of digital cameras has already created large markets for efficient digital image processing and compression techniques. Regardless of this, the great majority of current camera systems are still blind, lacking the ability to understand the contents of the imagery taken by the cameras. The goal of computer vision (machine vision) is to make useful decisions about real physical objects and scenes based on sensed images. The field is faced with a large number of scientific and engineering problems to be solved, and expertise from many disciplines is needed, including computer science, electrical engineering, mathematics, physics, and cognitive sciences.

The focus areas of the research are: 1) texture-based computer vision, 2) geometric image and video analysis, 3) machine vision for sensing and understanding human actions, and 4) vision systems engineering. Our approach of combining world-class basic research in chosen key areas with more applied research on vision systems engineering is quite unique among the research groups in computer vision.

The group and its members are active in the scientific community. For example, in 2004 Prof. Matti Pietikäinen served as an Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and Pattern Recognition journals; and members of the group were on the committees of several international conferences. Jari Hannuksela won the prize for the best M.Sc. thesis 2003 awarded by the Pattern Recognition Society of Finland for his Master’s thesis on facial feature based head tracking and pose estimation.

In November 2004, the group qualified for the second round of applications for the Finnish Centres of Excellence in the research programme 2006-2011 appointed by the Academy of Finland. The group is among 53 applicants that were selected from 143 plans of intent. It was also very significant that Prof. Janne Heikkilä was awarded quality unit funding from the University of Oulu for his project “Image analysis with geometric invariants and descriptors”.

Scientific Progress

Texture-Based Computer Vision

Image texture analysis is an important fundamental problem in machine vision. During the past few years, the group has developed theoretically and computationally simple, but very efficient nonparametric methodology for texture analysis based on Local Binary Patterns (LBP). The approach has evolved to represent a significant breakthrough, outperforming earlier methods in many applications. It is already widely used all over the world.

Recently, we proposed a novel facial representation for face recognition based on LBP features obtaining excellent results. In this approach, the face image is divided into several regions from which the LBP features are extracted and concatenated into an enhanced feature vector to be used as a face descriptor. However, it was unclear whether the high performance was due to the use of local regions (instead of an holistic approach) or to the discriminative power of LBP. We investigated this issue by comparing four different texture features in describing the appearance of local regions. The experimental results clearly showed the superiority of the LBP based approach.

Two approaches to facial expression recognition from static images were developed using LBP histograms computed over non-overlapping blocks for face description. In the first method, the Linear Programming
technique is adopted to classify seven facial expressions (anger, disgust, fear, happiness, sadness, surprise and neutral). In another approach, a coarse-to-fine classification scheme is used. Good results were obtained for the Japanese Female Facial Expression (JAFFE) database used in the experiments.

In addition to facial image analysis, we have started researching LBP based methods for other biometric identification tasks, including irises. The goal is to develop biometric identification algorithms for low-cost equipment such as mobile devices. The systems proposed so far require special instrumentation that limits their use. With the new methods, reliable identification would be possible with practically any camera.

A novel texture-based method for modeling the background and detecting moving objects from video sequences was developed. Each pixel is modeled as a group of adaptive local binary pattern histograms that are calculated over a circular region around the pixel. The method was evaluated against several video sequences including both indoor and outdoor scenes. It was shown to be tolerant to illumination variations, the multimodality of the background, and the introduction or removal of background objects. Furthermore, the method is capable for real-time processing.

Approaches to using LBP in content-based image retrieval were also studied. Block-based methods dividing the query and database images (or database images only) into blocks and comparing their LBP histograms were found to perform significantly better that methods based on global LBP histograms. The results for the LBP based approach were also better than those obtained with the widely used color correlogram features. Image databases taken from the Corel Image Gallery and from a German stamp database were used in experiments.

Research on visualization based learning in texture analysis was continued. Earlier, a visualization of LBP feature distributions with a self-organizing map was utilized in optical paper characterization and paper formation analysis. In the previous reporting period, the ISOMAP method was found to be promising in the visualization of high-dimensional texture feature data and in the training of a texture classifier. The research on utilizing the ISOMAP in texture analysis was continued. A visualization based approach was developed for finding the most discriminating LBP features and building a labelled training set for a texture classifier. The approach was utilized in paper texture analysis. A similar approach was also applied to wood grain texture analysis to study visually how local texture features can discriminate different grains.

The problems of texture image labelling were considered and a very promising labelling framework was developed. Textures were modeled with complementary measures including various versions of the LBP and Gabor features. Combined use of active learning, co-training, and visualization based learning was applied to feature data, enabling comprehensive, accurate, and user-friendly texture labelling. Also research concerning unsupervised learning of discriminative texture models from large data sets was initiated.

Research on the unsupervised learning and dimensionality reduction method, locally linear embedding (LLE), was continued. A new semi-supervised framework for visualizing high-dimensional data has been proposed. It is based on a combination of multiple machine learning algorithms, i.e. metric learning, intrinsic dimensionality estimation, linear feature extraction, and nonlinear feature extraction.

As a spin-off of the learning research mentioned above, the K-Local Hyperplane Distance Nearest Neighbor algorithm together with feature normalization was adapted for one of the important bioinformatics tasks - protein fold recognition. Fold recognition concerns structure (rather than sequence) of proteins and therefore it is very useful when there is a weak identity between protein sequences. Knowing protein structure often simplifies discovering protein function, which is an ultimate goal in biology and medicine. A comparison (on the publicly available dataset) with the state-of-the art in pattern recognition - support vector machines - demonstrated the superiority of the chosen approach and encouraged further exploration.

Geometric Image and Video Analysis

The research on geometric image and video analysis has been focused on two main topics: 1) geometric invariants and descriptors, and 2) tracking and motion estimation. In geometric invariants and descrip-
tors, a novel statistical approach has been developed to produce invariants that are insensitive to various geometric transformations. As a result, an image transform invariant to nonlinear perspective mapping has been proposed. The affine invariant multi-scale autoconvolution (MSA) transform introduced earlier is based on the same principle. Another affine invariant transform has also been developed, which requires only a fraction of the computation needed by the MSA, but its performance is superior to, for example, affine invariant moments. The research on image invariants has resulted in several scientific articles, and one of them has been accepted by the most respected journal in the field.

The group has discovered a means of estimating geometric transformation parameters using the same statistical approach as with the invariants, enabling us to apply those techniques also for image registration. Based on this solution, a registration method has been derived using the MSA transform that is capable of extracting affine transformation parameters between two images without requiring any point or feature correspondences. A further approach for featureless image registration has been developed that is based on the shift-invariance property of the image bispectrum. As only the phase information is utilized, the method is very robust against illumination changes, occlusions and background clutter. The method is capable of estimating the parameters of the similarity model including scale, rotation and translation.

In tracking and motion estimation, a new method for computing the statistical uncertainty of the block-motion estimates has been invented. Using this method, it becomes possible to estimate global camera motion quickly and reliably. The robust estimator developed tolerates also inconsistent motion regions in the scene. A video stabilization method utilizing this concept has been demonstrated with excellent results. The same approach has also been a basis for a new user interface technology for mobile terminals such as smart phones. Our solution utilizes a camera as a motion sensor that provides the motion parameters of the device. This information can be used for controlling the user interface so that, for example, browsing is performed simply by moving the terminal with respect to some static background. This solution has been implemented on a Symbian operating system, and demonstrated with a smart phone.

The motion estimation research has resulted in a new algorithm for computing the motion vectors for the H.264 video coding standard with various block sizes. This method is an extension of the multilevel successive elimination algorithm (MSEA) originally developed for fixed sized image blocks. The group has also investigated solutions for analyzing video and detecting moving persons based on block-motion vectors. A video surveillance system monitoring room entrance has been demonstrated using the ideas developed in this research. Enhancement of surveillance video quality is another topic of research carried out in the group during the last year. As a result, solutions for real-time software based video deinterlacing, and content-adaptive noise filtering were implemented. Methods for super-resolution image enhancement and motion deblurring were also investigated.

Machine Vision for Sensing and Understanding Human Actions

Machine vision provides a unique capability for understanding human actions. It is based on non-contact, passive sensing technology, and a large area can be covered by a single sensor. Our hypothesis is that it is possible to realize a large number of useful machine vision applications by using several distributed vision sensors that each adds a new view or viewing angle. A new framework of building such systems is being studied. An intelligent room equipped with multiple cooperating vision units is used as a test environment.

Human-computer interaction is one of the key issues in the development of proactive systems. Commun-
carnation with the proactive system can be based on smart user interfaces where the system derives the user input from, for example, gestures or facial expressions. In our group, we have investigated two types of approaches for camera based user interfaces. In the first approach, the system tracks the facial features of the user and determines the gaze direction, which can be directly used to control the pointer of the user interface in the same way as the mouse is used in a traditional PC user interface. The other approach developed is more suitable for a mobile system with an internal camera, and it is based on global motion estimation that produces the motion parameters of the device. These motion parameters can be exploited for controlling the user interface, which enables browsing and zooming large objects that do not fit into the display as such. Both approaches have been demonstrated with real-time implementations.

An experimental system for access control based on face recognition was under development. The system opens the door of a laboratory room only for selected persons while denying access to the others. It is implemented with novel methods developed by the group, including color based face detection using a skin locus model and hierarchical filtering, model selection for view based face recognition using the locally linear embedding algorithm and K-means clustering, and texture based face recognition using local binary patterns.

We have considered distance education as a practical problem that could greatly benefit from the technology under development. The classroom can be actually considered a smart room that has been equipped with sensors like microphones and cameras. An experimental system is being developed, which will assist the lecturer to select the most appropriate camera sources and visual angles that will be transmitted to the students. The system makes the decisions by analyzing the images obtained from the different cameras. Currently, the system has three cameras: an overall camera, a pan-tilt camera and a document camera.

The research made in co-operation with the French research center INRIA Rhone-Alpes addressed the problem of tracking motion of articulated objects from their 2-D silhouettes gathered with a small number of synchronized cameras. The method relies on (i) building 3-D observations (surface patches) from image silhouettes using a feed-forward method and (ii) on fitting an articulated object model to these observations through minimization. The objective function measuring the discrepancy between the model and data takes into account both the scaled algebraic distance from data points to model surface and the difference in orientation between observed surface patches and normals to the model’s surface. The results from the project are very promising. The new innovation, utilizing surface normals in model fitting, increases the method’s efficiency allowing more noisy input data.

**Vision Systems Engineering**

Vision systems engineering bridges the gap between algorithms and applications. It guides the methodological research to directions that improve the likelihood
of practical exploitation, finding and identifying approaches, architectures, methodologies, and algorithms that enable the building of useful real world systems. Solutions from low level image processing to even equipment installation and operating procedures are considered.

Visualization based non-supervised training methodology that is currently used in a number of industrial inspection systems on the market, from food sorting to lumber inspection. The central idea is to define the classification boundaries graphically to a 2-D projection of the multidimensional feature space, completely avoiding the tuning of numerical parameters. The non-supervised learning and dimensionality reduction techniques used include Isometric Mapping (ISOMAP), Locally Linear Embeddings (LLEs), and Self-Organizing Maps. The implemented systems employ non-segmenting image analysis techniques based on color and/or LBP histogram features.

In 2004, our group started a research project that targets the creation of a lumber strength grading methodology. This work is based on the use of LBP histogram features in combination with visualization based non-supervised training and FEM (Finite Element Method). We are also developing a solution for measuring the particle size distribution of a crushed aggregate used as the stone material in road pavements and concrete. The methodology is based on the non-segmenting approach, unsupervised classification methods, and texture features such as LBP. The goal is to find a means for implementing an on-line measurement instrument for an industrial environment. The non-supervised visualization based approach is used even for plant vitality measurement from spectrographic imagery. This work is at the research stage and aims at the utilization of photosynthesis and water content information in analyzer systems.

The dynamic ranges of the camera sensors are rather limited. As a cure we are endeavouring to create techniques that enable the capturing of good quality images from unevenly illuminated scenes. With a high frame rate camera this could be done by, for example, using alternating integration times for sequential frames. The proper integration times could be determined via content analysis, for example, by evaluating the quality of human “skin signal”. This image acquisition technology will be useful in proactive systems and visual surveillance applications.

We are also pursuing novel computer architectures to find efficient means for implementing image and video based systems. To minimize power dissipation, a problem for both performance computing and battery powered devices, hardware accelerated techniques need to be favored. Previously we have developed Number Theoretic Transform (NTT) based techniques that enable exploiting correlation and convolution theorems to cut the number of memory accesses, when compared to spatial domain computations. This work is being continued, while simultaneous multithreading techniques are studied to enable employing fine grained accelerators in an efficient manner, for instance, to implement real-time video codecs.

Exploitation of Results

The results of our research were applied to real-world problems in many projects, often in collaboration with industrial and other partners. Some examples of exploitation are described below.

In the wood inspection area, several board edgers using the visual training approach have been installed at sawmills by our industrial partner inX Systems. These have become fully operational within hours of installation unlike earlier systems which required at least a several weeks training period. Recently, inX Systems obtained the first ECVision (Excellence on Cognitive Vision Systems) Prize of the EU for its OptiGrader machine vision system which applies our research results.

The research in texture analysis and visualization methods has given birth to a new company, Intopii. The company is applying the results of the research to industrial visual inspection problems and is licensing its technology world-wide. Intopii currently employs four people full-time, and is growing steadily.

Our method for on-line paper characterization has also been on trial use in several paper mills in Finland and abroad. Our industrial partners have expressed interest in utilizing visualization based data analysis meth-
ods, like ISOMAP, in the training and further analysis of the paper texture data.

The research in video sequence processing and analysis has contributed to recent video codec products that have been integrated into mobile communications devices. Although the key role of the researchers was mostly in enabling rapid product development, the result is a convincing demonstration of the benefits of long term research.

**Future Goals**

We will continue to strengthen our long term research, researcher training and international activities. We will also continuously seek opportunities for exploitation of our research results by collaborating with partners from industry and other research institutions in national and international research programs and projects.

**Personnel**

| professors & doctors | 8 |
| graduate students    | 15 |
| others               | 11 |
| **total**            | **34** |
| person years         | 25 |

**External Funding**

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>142 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>176 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>232 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>209 000</td>
</tr>
<tr>
<td>EU + other international</td>
<td>4 500</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>763 500</strong></td>
</tr>
</tbody>
</table>

**Selected Publications**


Background and Mission

The Machine Vision Group (MVG) of the University of Oulu, established in 1981, is the leading group of its kind in Finland. It has achieved a highly respected position in its field and its results have been widely exploited in industry. The objective of the group is to carry out leading edge long term research on machine vision technology. In practice this will be achieved through top-class fundamental research in chosen focus areas, close interaction between basic and applied research, and in-depth international collaboration.

Vision is a key component for building artificial systems that can perceive and understand their environment. Currently, the majority of camera based systems are cognitively blind, lacking the capability to understand the contents of the imagery. The goal of computer vision is to find methods that provide this capability, while machine vision applies those methods into practice, combining computer, optical and automation engineering.

In 2005, the focus areas of the research were: 1) texture based computer vision, 2) geometric image and video analysis, 3) machine vision for sensing and understanding human actions, 4) learning in machine vision, and 5) vision systems engineering.

The year 2005 meant a significant growth and deepening in the international activities of the group. We have established active collaboration with some of the world’s leading institutions and top scientists. We have had in-depth collaboration with the University of Maryland (USA) since the early 1980’s, and more recent partners include the Chinese Academy of Sciences and INRIA Rhone-Alpes (France). We also carry out joint research efforts with the Academy of Sciences of the Czech Republic and the University of Freiburg (Germany).

We foster researcher mobility to and from our unit. Five of our researchers made research visits to our partner institutions during the reporting period. We also have visiting postdoctoral researchers from abroad. This mobility has led to joint research projects and co-authored publications.

The group and its members are active in the scientific community. In 2005, Prof. Pietikäinen served as an Associate Editor of the Pattern Recognition journal and finished his two term service as the Associate Editor of the IEEE Transactions on Analysis and Machine Intelligence journal (IEEE TPAMI). The professors of the group were program committee members of several international conferences and many researchers of the group served as reviewers for various journal and conference articles. Members of the MVG were increasingly successful in producing high-quality publications. For example, three papers were accepted for the IEEE TPAMI journal, which is currently ranked #1 among all 209 electrical engineering titles in the Journal Citation Report.

The Machine Vision Group qualified for the second round of applications for the Finnish Centre of Excellence in the research program 2006-2011, obtaining outstanding evaluations. It is also significant that Prof. Heikkilä has received “quality project unit” funding from the University of Oulu for the period 2004-2006.

Scientific Progress

Texture Based Computer Vision

Image texture analysis is an important fundamental problem in machine vision. During the past few years, the group has developed theoretically and computationally simple, but very efficient nonparametric methodology for texture analysis based on Local Binary Patterns (LBP). The approach has evolved to represent a significant breakthrough, outperforming earlier methods in many applications. It is already widely used all over the world. In 2005, our research focused on robust LBP based texture descriptors, texture based methods in facial image analysis, and analysis of motion and dynamic textures.

Recently, we have begun research on texture based region descriptors which are robust with respect to severe image transformations, including affine geometric transformations, illumination changes, blurring and noise. Two extensions to the local binary pattern operator were developed. The first one is a filter based approach which shows that the LBP operator can be seen as a special filter based operator. This view of LBP can be useful in analyzing the properties of the operator and in designing image descriptors for spe-
pecific tasks. Another proposed extension is using soft histograms which were shown to enhance the performance of the operator for noisy images.

Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. In 2005, we began to investigate SIFT-like descriptors which are based on local binary patterns, with very promising preliminary results. This research is carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhone-Alpes, France.

In 2004, we proposed a novel facial representation for face recognition based on LBP features, obtaining excellent results. In this approach, the face image is divided into several regions from which the LBP features are extracted and concatenated into an enhanced feature vector to be used as a face descriptor. Our approach has evolved to be a growing success. It has been adopted and further developed by many research groups.

Recently, we began to investigate texture based methods for face recognition which are robust against factors such as illumination changes, blurring, occlusion, and noise. For test image sets we have been using Face Recognition Grand Challenge problems as recently defined by the US government (http://www.frvt.org/FRGC/). In this research, we collaborate with Prof. Rama Chellappa from the University of Maryland, USA.

Research on real-time face detection was begun with an aim to develop methods suitable for real-world environments such as video surveillance and mobile devices. Our face detection method, based on a compact LBP based feature space and support vector machine classification, introduced at CVPR 2004, was modified to use a cascade structure and a fast classification scheme based on FloatBoost learning. A real-time operation was achieved with as good detection accuracy as the original, which was a much slower approach. Further, the system is extended to deal with multi-view face detections yielding encouraging preliminary results. The implementation of these detection algorithms on mobile phones under the Symbian platform has also begun.

Color is a very powerful and useful cue for face detection, but unfortunately it is also sensitive to changes in illumination and other factors which are easily encountered in unconstrained environments. Considering these facts, a new face detection approach has been developed in 2005. The proposed approach consists of pre-processing the images with skin locus, a method for skin color modeling developed earlier in our group, and then performing an exhaustive search in and around the detected skin regions. The exhaustive search is performed by using a new two-stage SVM based approach, exploiting the discriminative power of the LBP features. The results obtained are interesting in the sense that the proposed approach inherits the speed from the color based methods, and the efficiency from the gray scale based methods. Further, an investigation on the use of color information to support the detection of partially occluded faces has begun.

In addition to the facial image analysis, we have also launched an investigation on texture-based methods for iris recognition. Our goal is to develop algorithms for low-cost equipment such as mobile devices. As the first step, a fast iterative algorithm was developed for accurate localization of irises in eye images. In experiments, the algorithm gave almost faultless results for the vast majority of the images in the CASIA database and converged into a solution in a few iterations.

A real-time method for modeling the background and detecting moving objects from video sequences has been developed. Each pixel is modeled as a group of adaptive local binary pattern histograms that are calculated over a circular region around the pixel. A paper written on the topic was accepted for the IEEE Trans. Pattern Analysis and Machine Intelligence journal. The original method required a non-moving camera, but recently we have developed an extension for moving cameras by combining it with a robust feature-based image mosaicing method. Image mosaicing is the process of stitching together two or more images of the same scene taken from different viewpoints or viewing directions. Our method utilizes the SIFT feature detector that has thus far been rarely utilized in image mosaicking tasks.

Image mosaicing from single sequences.
We have also begun to study dynamic textures (i.e. textures in motion) which are important, for example, in video retrieval applications. Examples of dynamic textures are sea-waves, smoke, foliage, fire, shower, whirlwind, and a talking face. A volume LBP operator was developed which combines the temporal and spatial information (i.e. motion and appearance) together. A simplification of the method based on concatenated LBP histograms computed from three orthonormal planes was also studied. Our approach has outperformed the state-of-the-art methods (e.g. those based on optic flow) in experiments with the MIT and DynTex databases.

Geometric Image and Video Analysis

The research on geometric image and video analysis has been focused on the following topics: 1) geometric invariants, 2) image registration, 3) geometric camera calibration, 4) motion analysis and video stitching, 5) video coding, and 6) image processing for an MRI compatible robot.

In geometric invariants, we have continued working with global affine invariants. We have recently started collaboration with Prof. Hans Burkhardt from the University of Freiburg, Germany, which is related to object recognition using geometric invariants. We have developed a framework for constructing invariants based on multiscale image representations. This framework includes multiscale autoconvolution (MSA) that was developed in our group earlier, but it also includes two other novel methods proposed more recently. The first one is called the spatial multiscale affine (SMA) invariant, which is a computationally inexpensive approach for deriving affine invariant descriptors. The second method is based on ridgelets. It has more degrees of freedom, and its discrimination power seems to be better than with the other methods. We have also been able to combine affine moment invariants and multiscale invariants. Experiments have indicated an increased performance of the combined method. This work has been carried out in co-operation with Prof. Jan Flusser from the Academy of Sciences of the Czech Republic.

Our research on image registration was inspired by the probabilistic framework behind MSA. A central part of the research was to extend the methodology of computing invariants to registration. This was done by deriving such image features that are not invariant but instead allow the recovery of affine transformation parameters between two images. In particular, two different novel registration methods were proposed, which are both based on recent techniques for generating affine invariants. The methods directly utilize the intensity distribution of the images and no separate feature extraction or correspondence search is required. In addition, the proposed methods can be extended to the registration of point sets, which can be either two or three-dimensional. Since the proposed registration techniques are non-iterative they might be particularly useful as a means for obtaining an initial registration that is needed for iterative registration methods, such as the iterated closest point algorithm, for example.

In geometric camera calibration, the aim of the research was to develop an accurate camera calibration method that would be suitable for lenses with a very wide field of view, such as fish-eye lenses which may have a field of view over 180 degrees. Fish-eye lenses are convenient in such applications where a very wide angle of view is needed but their use for measurement purposes has been limited by the lack of an accurate, generic, and easy-to-use calibration procedure. We have therefore proposed a generic camera model which is suitable for fish-eye lens cameras as well as for conventional and wide-angle lens cameras, and a calibration method for estimating the parameters of the model. The calibration method is based on viewing a planar object which contains control points in known positions. In particular, during the year 2005, the earlier work in camera calibration was further developed in order to make the calibration method work for lenses that have a field of view significantly over 180 degrees. The results of this research, which was begun at the Helsinki University of Technology, have been implemented into a Matlab toolbox that is available to the public.

In our research on motion analysis, we have investigated techniques for estimating the confidence of block motion candidates. We have improved our previous solution that thresholds motion profiles and determines...
uncertainty covariance matrices for the corresponding motion vectors. The information obtained provides a mechanism for estimating global camera motion from various kinds of features including lines and edges. Two applications have been considered: video stabilization and motion control of a mobile terminal. Both applications have been implemented on a Symbian based smart phone. In our recent work, we have developed a new likelihood measure that enables fast evaluation of arbitrary motion candidates based on a statistical model. This solution can be used for extracting several motion regions from image sequences when augmented with a particle filter.

In video stitching, the aim is to generate panoramic images from video sequences. In our group, we have developed an algorithm that locates moving objects in the sequence and constructs the mosaic so that the moving objects are shown correctly in the panorama image. The algorithm is also capable of correcting the lighting differences between frames in order to fade out the seams from the result. A new approach in our solution is to select the frames for the mosaic based on their quality. The quality is measured with a few parameters, like the amount of blur in the image. Special attention has also been paid to the computational efficiency of the algorithm.

A panorama image from a video sequence.

In video coding research, the group has devised a fast method for distributing bits optimally over successive video frames, which improves the average image quality by up to a decibel, without affecting the encoded video bit rate. The new method works best with the advanced coding options available in the H.264 video coding standard, but is applicable to older standards as well. In particular, the method is completely standards compliant. New block motion estimation methods for video coding have also been investigated. The objective is to reduce the computational complexity of the motion estimation step without causing significant degradation in the image quality or compression efficiency. Some promising results with the H.264 video coder have been achieved in the preliminary experiments.

In 2005, our group started a new research project funded by the Academy of Finland, where the main target is to develop image processing capabilities for a surgical robot operating in a magnetic resonance imaging (MRI) environment. Due to the strong magnetic field, the robot should be made of some MRI compatible materials instead of conventional ferromagnetic materials. The methods and technical solutions under development are expected to improve the usability of the robot, and make it possible to perform operations where the presence of medical staff in the immediate vicinity of the patient is unnecessary. This also reduces the number of errors and distortions occurring in MRI pictures. The project belongs to the Information Technology in Mechanical and Automatic Engineering Research Programme (KITARA) of the Academy of Finland, and there are four parties and separate projects in the consortium.

**Machine Vision for Sensing and Understanding Human Actions**

Machine vision provides a unique capability for understanding human actions. It is based on a non-contact, passive sensing technology by which a large observation area can be covered by a single sensor. Our hypothesis is that it is possible to realize a large number of useful machine vision applications by using several distributed vision sensors each of which add a new view or a viewing angle. A new framework for building such systems is being studied. An intelligent room equipped with multiple cooperating vision units is used as a test environment.

Human-computer interaction is one of the key issues in the development of proactive systems. Communication with the proactive system can be based on smart user interfaces where the system derives the user input from, for example, gestures or facial expressions. In our group, we have investigated different approaches to vision based user interfaces. One novel approach is to control the user interface of a handheld device based on motion information measured from images that are captured by a built-in camera. We have developed an efficient solution for estimating the camera motion from arbitrary features. This solution utilizes our motion analysis framework, where each feature is weighted by its confidence. We have implemented the motion estimation module on a smart phone, and used it for different applications. One ap-
Application is browsing and zooming of a large bitmap on the display using hand movements. The second application is a drawing tool, which makes it possible to create simple line drawings by moving the device.

We have also developed a fast method for recognizing handwritten letters and numerals from the strokes created with this tool. The recognition accuracy achieved is around 95%.

Another approach investigated in our group for human-computer interaction is based on hand tracking. We have developed a real-time system for tracking the user’s hand position and shape from a camera located above the display. The trajectories obtained are used for moving the cursor on the screen, and the shape information, including the number of the fingers visible for giving other commands.

Videoconferencing is also a potential application area of the camera based user interface technology. A special case of video conferencing is distance education, where the teacher needs to switch constantly the sources of the video feed between different cameras. For the teacher, this is annoying, and in practice these actions may easily be forgotten, which may have a detrimental effect on teaching and learning. The usability of the system could be significantly improved by changing the sources automatically on the basis of the teacher’s gestures and behavior. In our research, we have developed an automated system that takes care of both the selection and switching of the video source in a distance education situation in a real lecture room environment. The stimulus for changing the source is obtained directly from the video cameras, and no other sensors are needed. For instance, if the teacher uses the document camera, the system recognizes this event, and, as a response, it changes the source of the video feed to the document camera.

As a part of the Tekes funded PersonID project, different methods for visual surveillance were investigated. A number of background segmentation and tracking algorithms, together with novel applications from both the commercial and research angle, were collected into the form of a literature survey. While going through the current state of development, a real-time testing platform was implemented to evaluate the existing and new potential approaches. Some initial demonstrations were given to the companies participating in the project. The main research interest was on methods that are well suited for person detection and tracking in surveillance video. Online feature selection was combined with a MeanShift tracker to track people, and also a CamShift-based approach was tested. Current development focuses on tracking of multiple persons using texture and color features.

Ways of describing human activities were studied from two different points of view. Both the 2D and 3D approaches were explored. Firstly, a silhouette based 2D method was developed, in which affine invariant Fourier descriptors are used to describe the human pose in a frame. A support vector machine classifier is used for recognizing the posture class and Hidden Markov Models are used for classifying posture sequences. More recently, a novel 3D approach was developed in co-operation with Prof. Larry S. Davis from the University of Maryland. The new method is inspired by a model of human motion found in psychological studies. The method is capable of modeling and recognizing natural human movements, such as reaching and striking, which have highly variable target locations and thus huge variation in the actor’s body configuration. The method has been tested on 3D motion capture data with promising results.

The research made in co-operation with Dr. Radu Horaud, INRIA Rhone-Alpes, was continued, addressing the problem of tracking motion of articulated objects from their 2D silhouettes gathered with a small number of synchronized cameras. The method relies on building 3D observations (surface patches) from image silhouettes using a feed-forward method and on fitting an articulated object model to these observations through minimization. The objective function measuring the discrepancy between the model and data takes into account both the scaled algebraic distance from data points to the model surface and the difference in orientation between observed surface patches and normals to the model’s surface. The results from the project are very promising. The results were published in a conference article and a journal article has been prepared for submission.
Research on video-based gait recognition, carried out earlier by one of our postdoctoral researchers, Guoying Zhao, at the Chinese Academy of Sciences, was completed. Gait recognition is one of the behavioral biometric methods, used to signify the identity of individuals in image sequences ‘by the way they walk’. From a surveillance perspective, gait recognition is an attractive modality because it may be performed surreptitiously at a distance. On the basis of 2D analysis, a method based on fractal scale and wavelet moments was proposed to describe and automatically recognize gait. Good results were obtained for the USF database used in experiments. More recently, 3D gait recognition was also investigated. A 3D human model is set up, and the walking is tracked by applying a local optimization algorithm. Static and dynamic features are extracted for recognition. The proposed 3D approach is robust to occlusion and changes of viewpoint. Moreover, better results than with 2D analysis are achieved for sequences containing difficult surface variations.

Learning in Machine Vision

The ability to learn plays a critical role in solving many machine vision problems, and represents one of the major challenges for research. Most of our recent research on learning is related to manifold learning and nonlinear dimensionality reduction. It is well known that high-dimensional feature data often lies in a lower dimensional manifold in the observation space. We have been studying methods for learning these low dimensional manifolds from image or video data. The low dimensional manifold is then used, for example, for visualizing data, or selecting appearance models or features for recognition.

Research on the unsupervised learning and dimensionality reduction method called locally linear embedding (LLE) was continued. A problem with basic LLE is that it operates in a batch mode, i.e. when new data arrives, one needs to rerun the entire algorithm with the original data augmented by the new samples. We proposed a novel incremental version of LLE to solve this problem, and experimentally demonstrated its usefulness in various problems.

Research on visualization based learning in texture analysis was continued. An Isomap based method was used to analyze and visualize paper and wood grain textures. Our paper on lumber grain analysis received an award for the best paper at the International Conference on Quality Control by Artificial Vision (QCAV 2005). The method was found to be useful in early stage data analysis when a texture classifier is constructed and trained. A framework for texture image labeling was developed utilizing the combined use of active learning, co-training, and visualization based learning. With this framework, human work in the texture sample labeling is reduced because large amounts of samples are automatically processed and confidently labeled. In addition, research was started on learning and using contextual constraints and information in texture based outdoor scene image labeling. The initial results show improvements compared to the local classifiers that were used earlier in our research.

Interdisciplinary research in using manifold learning in brain activity analysis which utilizes functional magnetic resonance imagery (fMRI) was initiated together with the Department of Radiology (the largest unit in the Nordic countries) at the University of Oulu, and the Chinese Academy of Sciences (Prof. Yu-Feng Zang, National Laboratory of Pattern Recognition). Our goal is to develop clinically applicable diagnostic methods modeled according to the solutions originally developed for learning in visual inspection applications. The early detection of neurodegenerative diseases, such as Parkinson’s and Alzheimer’s, is what we are aiming for. Resting state fMRI (functional Magnetic Resonance Imaging) is used for obtaining the data. A project proposal on this topic was accepted for the NEURO 2006-2009 program of the Academy of Finland.

Learning representations of objects or object categories from a set of training images for the recognition of previously unseen object instances has recently become a popular field of research in computer vision. Our research in this field has just begun in cooperation with INRIA Rhone-Alpes. The research has concentrated on development of methods for the clustering and classification of high-dimensional visual data. The visual descriptors commonly used in object recognition are often high-dimensional and this penalizes classification methods and consequently recognition. A new approach, proposed by the researchers at INRIA, combines the dimension reduction and classification in a unique step. This is achieved by using a new kind of Gaussian mixture model which identifies the specific subspace of each class. In object recognition, the new clustering approach is applied for local image descriptors and embedded into a probabilistic recognition framework.
As a spin-off of our manifold learning research, the K-Local Hyperplane Distance Nearest Neighbor algorithm, together with feature normalization, was adapted for one of the important bioinformatics tasks - protein fold recognition. Fold recognition concerns the structure (rather than a sequence) of proteins and therefore it is very useful when there exists a weak identity between protein sequences. Knowledge of the protein structure often simplifies discovering protein function, which is an ultimate goal in biology and medicine. A comparison (on the publicly available dataset) with the state-of-the art in pattern recognition - support vector machines - demonstrated the superiority of the chosen approach and encouraged further exploration. Research on these topics has involved collaboration with scientists from the University of Porto, Portugal and Tallinn University of Technology, Estonia.

**Vision Systems Engineering**

To enable useful real-world systems, our vision system engineering research provides guidelines for methodological research, helping to identify attractive approaches, architectures, and algorithms, as general purpose computing is seldom a realistic option. In practice, solutions from low-level image processing to even equipment installation and operating procedures need to be considered simultaneously. The roots of this expertise are in our industrial visual inspection research in which we met extreme computational requirements already in early 1980’s. We have participated in designing several industrial systems in this area. The field is still very active, and our collaborative vision computing architecture research is a recent spin-off.

Machine vision applications are characterized by both high data input rates and computational costs. For instance, a typical raw digital video rate is around 10 Mpixels/s and its processing requires at least a few hundred operations per pixel, requiring multiple GOPS of computing power. Often this needs to be done in a small package, such as a mobile communications device that may allocate at most 500 mW for application processing like video coding or person identification. This prevents implementations based on conventional processors even in the future, but hardware acceleration is needed. Currently, only monolithic long latency accelerators improve energy efficiency, but they are rigid and costly to design and verify.

Together with the Technical Universities of Tampere (Prof. Jarmo Takala) and Helsinki (Prof. Petri Vuorimaa), and the Åbo Akademi (Prof. Johan Lilius) we are concentrating on improving the energy-efficiency of embedded high performance computing. The desired system solutions exhibit fine grained adaptable hardware accelerators to improve the silicon area efficiency via run-time re-use, while simultaneous multithreading is favored in hardware/software interfacing to reduce software overheads instead of an interrupt mechanism. We are essentially targeting a new paradigm for embedded computing. The first step in the research is the redesign of an MPEG-4 hardware encoder to provide a reference point on the relative power of the approach. Our earlier related research includes the development of Number Theoretic Transform (NTT) based techniques that enable exploiting correlation and convolution theorems to cut the number of memory accesses when compared to spatial domain computations.

Visual inspection is economically still the most important application area of machine vision. Inspection systems can be relatively expensive, as long as they provide high added value, and are therefore attractive testing grounds for new technologies. Typical inspection targets include part assemblies in the electronics and car industry, continuous webs such as paper, steel and fabrics, and natural materials such as wooden boards and coffee beans. Many of these targets are textured and colored, such as wood, and the inspection problem is solved best with target specific methods.

Several inspection systems based on our results are now marketed by our industrial partners for applications ranging from coffee bean sorting to paper formation measurement. Recent systems employ both our visual training approach and LBP based texture analysis methods. Currently, we are investigating methods and means for building visual inspection systems for exceptionally demanding applications such as non-destructive, non-contact dynamic strength grading of wooden boards. The underlying observation is that the strength is different to the axial, tangential and radial directions of the wood grain, while the behavior of the grain is affected differently by sound and dry knots. After color and texture based image analysis has provided the grain and knot information, a 3D Finite Element Model (FEM) is built and analyzed for strength.
Real-time operation requires super-computing speeds (> 100 GFLOPS) from a reasonably priced embedded system and this represents a big challenge.

An example of a three-dimensional FEM model of the longitudinal stress field of a board.

The dynamic ranges of the camera sensors are rather limited. As a cure, we have endeavored to create techniques that enable the capturing of good quality images from unevenly illuminated scenes. With a high frame rate camera this could be done by, for example, using alternating integration times for sequential frames. The proper integration times could be determined via content analysis, for example, by evaluating the quality of a human “skin signal”. Our good preliminary results suggest that this kind of image acquisition technology could be useful in proactive systems and visual surveillance applications, for example.

**Exploitation of Results**

Our approach of combining world-class basic research with more applied research on vision systems and systems engineering is quite unique, giving rise to our research having a great practical impact. We conceive of machine vision research as a remarkable field of science that contributes to the competitiveness of Finnish enterprises by developing methods and techniques for improving the performance and usability of industrial machines and products.

In 2004 our industrial partner, inX Systems, obtained the first ECVision (Excellence on Cognitive Vision Systems) Prize from the EU for its OptiGrader machine vision system which applies our research results in visualization-based wood inspection system training. Several board edgers using the visual training approach have been installed in sawmills. These have become fully operational within hours of installation, unlike earlier systems which required a training period lasting at least several weeks.

Another example of the impact of our work is that in 2005 Intopii Oy, a spin-off company of our texture research, entered into a cooperative agreement with the Cognex Corporation, the world’s leading supplier of machine vision systems.

The research in video sequence processing and analysis has contributed to recent video codec products that have been integrated into mobile communications devices. Although the key role of the researchers was mostly in enabling rapid product development, the result is a convincing demonstration of the benefits of a long term research.

Our research on video filtering has yielded computationally efficient solutions for content dependent video enhancement and deinterlacing. These results have been utilized by one of our industrial partners, Mirasys; a company specialized in digital video recorders for visual surveillance.

Distance education has become an important part of our educational system, and it is also a rapidly growing area of business where different vendors provide technology and platforms for schools and universities. Our contribution in this field is a Distance Education Assistant, which is an automated system for switching video sources based on image analysis. This solution has already been adopted by Videra, the leading Finnish developer and applier of remote work technology.

**Future Goals**

The progress of science and technology has always been our best impetus. The research of Machine Vision Group has many prospects. Its innovations in camera based systems with image understanding capabilities provide great potential for new applications in areas such as proactive and ubiquitous computing, biometrics, visual surveillance, human-computer interfaces, multimedia, content based retrieval from databases and vehicle guidance. The results of the research are increasingly being published in top scientific journals as well as being made publicly known in the media. The group will continue its high-grade research and attempt to pursue a more stable financial foundation for its operation. We will also continue working closely with our international partners, industry and other interest groups. As always, we will foster the career development of our researchers and continue taking active part in the operation of Infotech and GETA Graduate Schools.

**Personnel**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>8</td>
</tr>
<tr>
<td>graduate students</td>
<td>18</td>
</tr>
<tr>
<td>others</td>
<td>10</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>36</strong></td>
</tr>
<tr>
<td>person years</td>
<td>26</td>
</tr>
</tbody>
</table>
External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>321 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>221 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>196 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>183 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>921 000</strong></td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


Background and Mission

The Machine Vision Group (MVG) is renowned world-wide for its expertise in computer vision. It works as a single well-focused research group and collaborates with other groups of similar status in Europe, USA and China. The research areas of the group range from generic computer vision methodologies to machine vision applications and vision systems engineering. The results of its research have been widely exploited in industry. The current areas of interest include face analysis, camera-based user interfaces, visual surveillance, smart environments and energy-efficient architectures for vision computing.

The MVG celebrated its 25th Anniversary in 2006. Highlights of the research as well as a short history of the group were published in a book entitled “From algorithms to vision systems – Machine Vision Group 25 years”. In September, the group organized an anniversary seminar with speakers from academic and industrial circles.

The group increased its visibility also by having a joint exhibition stand with the Intelligent Systems Group in the European Information Society Technologies IST 2006 event held in Helsinki in November.

The group has hosted several visitors from abroad, including from the Chinese Academy of Sciences (Prof. Stan Z. Li, Yu-Feng Zang, Jianguo Han and Rongkai Lu), University of Maryland (Prof. Larry S. Davis and Shuvra S. Bhattacharyya), and the Academy of Sciences of the Czech Republic (Prof. Jan Flusser). In addition, several domestic visitors, for example from the Ministry of Education of Finland and the Academy of Finland, were briefed about our research activities.

In 2006, the focus areas of the research consisted of: 1) texture-based computer vision, 2) geometric image and video analysis 3) machine vision for sensing and understanding human actions, 4) learning in machine vision, and 5) vision systems engineering, including new paradigms for embedded systems.

The group has established active collaboration with some of the world’s leading institutions and top scientists. It has had in-depth collaboration with the University of Maryland (USA) since the early 1980’s, and more recent partners include the Chinese Academy of Sciences and INRIA Rhone-Alpes (France). Joint research efforts with the Academy of Sciences of the Czech Republic and the University of Freiburg (Germany) are also being made.

The group fosters researcher mobility to and from our unit. Five of our researchers made research visits to partner institutions during the reporting period. The group also had visiting postdoctoral researchers and graduate students from abroad. This mobility has led to joint research projects and co-authored publications.

The group and its members are active in the scientific community. For example, in 2006-2007 Prof. Pietikäinen has served as an area chair of the top-ranking IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007). He is also a Co-Chair of Workshops for the International Conference on Pattern Recognition (ICPR 2008). The professors of the group were committee members of several other international conferences and many researchers of the group served as reviewers for various journal and conference articles. Prof. Matti Pietikäinen, Dr. Abdenour Hadid and Dr. Guoying Zhao lectured for a tutorial on computer vision using local binary patterns at the 18th International Conference on Pattern Recognition in Hong Kong.

In 2006, the Pattern Recognition Society of Finland elected Prof. Janne Heikkilä as its president. Esa Rahtu was chosen to serve as the secretary of the society and Dr. Sami Brandt as its treasurer.

The publication record of the group has improved both in quality and quantity. For example, since 2005 as many as six papers of the group have been accepted for the IEEE Transactions on Pattern Analysis and Machine Intelligence journal (IEEE TPAMI) which is the leading journal of the field and ranked among the top journals in electrical engineering and computer science in the Journal Citation Report.
The group and its projects have received excellent evaluations from several panels of international experts, for example in the recent evaluation of the Academy of Finland’s PROACT program.

**Scientific Progress**

**Texture-Based Computer Vision**

Texture is a fundamental property of surfaces. It can be seen almost anywhere. We have developed a novel methodology based on Local Binary Patterns (LBP), which has evolved to present a major breakthrough in texture analysis. It is already widely used all over the world. Our recent results suggest that the approach could offer significant potential for many important tasks in computer vision which have not been earlier regarded as texture problems.

In 2006, our research focused on robust LBP-based texture descriptors, contextual analysis of textured images, texture-based methods in facial image analysis, and analysis of dynamic textures. Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. A novel Center-Symmetric-LBP (CS-LBP) interest region descriptor which combines the strengths of the SIFT descriptor and the LBP texture operator was proposed. In extensive experiments, our descriptor performed better than SIFT for most of the test cases and about equally well for the remaining ones. Especially, the tolerance of our descriptor to illumination changes was clearly demonstrated. This research was carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhone-Alpes, France.

The fundamental properties of the Local Binary Pattern operator were analyzed in a study that aimed at making the operator more robust to noise. It was shown that the operator can be seen as a combination of derivative filters and a vector quantizer based on thresholding a set of filtered images. Soft histograms that make the operator continuous and robust for LBP, were introduced for LBP and filter based texture operators. It was shown that when the images to be analyzed have a high amount of noise, a significant increase in performance can be obtained by utilizing the soft histograms and choosing a filter set that is more robust with respect to noise.

Research on texture based outdoor scene image analysis was continued. A very efficient classification system combining local and contextual image texture information was proposed. Texture was modeled with LBP and classified locally with SVM. Another CRF-based classifier was used to improve the results by taking into account contextual image constraints.

In 2004, we proposed a novel facial representation based on LBP features, obtaining excellent results. A paper on this topic was recently published in IEEE Transactions on Pattern Analysis and Machine Intelligence. Our approach has evolved to be a growing success. It has been adopted and further developed by many research groups.

In the original version, the facial image is divided into small rectangular blocks and LBP histograms are computed within each of them. A similar representation has been adopted by most researchers that utilize LBP for face image analysis. Recently, we began a study on different representations for LBP coded face images that would be better suited for face images. Kernel based density estimation has served as a starting point for this methodology. The preliminary results using the challenging Face Recognition Grand Challenge database are promising and attest to the suitability of this approach.

Research on real-time face detection using local binary patterns was continued by conducting research on multi-view face detection, with an aim to develop a real-time face detector suitable for applications such as video surveillance, mobile devices and content based video retrieval. The goal was achieved using cascade classifiers with additional training samples. The new multi-view system uses LBP features in a coarse-to-fine detection strategy (pyramid architecture) embedded in a fast classification scheme based on AdaBoost learning.

Inspired by the works of Viola and Jones on the use of Haar-like features with integral images and that of Heusch et al. on the use of LBP as a preprocessing step for handling illumination changes, a robust approach for eye detection using Haar-like features extracted from LBP images was developed. In the system, the images are first filtered by an LBP operator and then Haar-like features are extracted and used with AdaBoost for building a cascade of classifiers. Very promising results were obtained in the experiments.

**Eye detection.**

In 2005, we began to study methods for analyzing dynamic textures, i.e. textures in motion. A volume LBP operator (VLBP) was developed which combines temporal and spatial information (i.e. motion and appearance). A simplified method based on concatenated LBP histograms computed from three orthogonal planes (LBP-TOP) was also proposed.

Recently, we extended these spatiotemporal descriptors to deal with non-traditional dynamic textures, such as facial expressions, in which local information and its spatial locations should also be taken into account. A representation which consists of dividing the face image into several overlapping blocks was introduced. The LBP-TOP (or VLBP)
histograms in each block are computed and concatenated into a single histogram. All features extracted from each block volume are connected to represent the appearance and motion of the facial expression sequence. This approach does not require error prone segmentation of lips and other facial features and it is robust against monotonic gray scale changes caused, for example, by illumination and skin color variations, and errors in face alignment. A recognition rate of 96.26% was achieved for recognizing six prototypic emotions, anger, disgust, fear, joy, sadness, and surprise, from the widely used Cohn-Kanade database, even though we used in our experiments simpler image preprocessing and a larger number of people (97) and sequences (374) than most of the others have used. In January 2007, a paper on this topic was accepted for IEEE Transactions on Pattern Analysis and Machine Intelligence.

A spatiotemporal LBP approach was also adopted for recognizing faces in videos during the three month research visit of Dr. Hadid to Prof. Stan Li’s group at Chinese Academy of Sciences. To tackle the problem of selecting only the spatiotemporal information which is useful for recognition, the AdaBoost learning technique was used. The idea was to classify the facial information into intra and extra classes, and then use only the extra-class LBP features for recognition. Very promising preliminary results were obtained.

A 3D model of a sewer pipe reconstructed from a video sequence scanned by a robot moving inside the pipe.

We also began to investigate recognition of isolated phrases using only visual information. A new appearance feature representation based on LBP-TOP descriptors was proposed, taking into account the motion of the mouth region and time order in pronunciation. Positions of the eyes determined by a robust face and eye detector are used for localizing the mouth regions in face images. Spatiotemporal local binary patterns extracted from these regions are applied for describing phrase sequences. In our experiments, promising accuracies of 62% and 70% were obtained in speaker-independent and speaker-dependent recognition, respectively.

**Geometric Image and Video Analysis**

Imaging geometry provides the basic mathematical tools for analyzing the spatial properties of images. For image sequences, motion is another important feature that describes the temporal changes between successive images. Topics dealing with both imaging geometry and motion have been explored in this part of the research.

During 2006 we have continued our research related to the structure-from-motion (SFM) problem. We have considered both the theory of geometric computer vision from statistical view points as well as SFM applications. For instance, one topic in our research has been the development of generic camera models which would allow more accurate modelling of real cameras than the previous models, and which would be suitable for both omnidirectional and conventional cameras. We have also investigated structure-from-motion techniques for long image sequences taken with a fish-eye lens camera. Related to this application, a robust method for modelling tubular structures from a sparse set of 3D points has been developed.
Pose estimation based on point correspondences is an essential problem in many computer and robot vision applications. Within this topic, we have studied evolutionary computation techniques for pose estimation of a known object from a single image. A method using only a few point correspondences has been developed for those cases in which traditional analytical and iterative pose estimation approaches often fail. The method does not need an accurate initial guess, it finds point correspondences automatically and it has been proved to be robust and reliable under various conditions.

We have also continued our research on geometric invariants, and maybe the most important individual achievement in this field was the development of a multiscale framework. This now allows us to neatly bind together all the affine invariant methods we have previously proposed. The new ideas in the framework also provide a general approach for dealing with geometric distortions in the images when constructing recognition or registration methods.

In many situations when the camera is moving, the images obtained become more or less blurred and this may cause some problems for image analysis methods. This was the starting point for our research on blur invariant pattern recognition and image registration. As a result, we have discovered new blur invariant features based on the phase-only spectrum of the image. These features are invariant to centrally symmetric blur, such as linear motion or out-of-focus blur. Based on this research, we developed a method called blur invariant phase correlation (BIPC) for registration of blurred and translated images. These new features were also used for object recognition in blurred images. To achieve translation invariance, we used bispectra instead of the ordinary spectra.

In video coding, we have developed a fast search algorithm called Rotation Search, which is a general optimization method suitable for optimizing very high dimensional discrete functions. We implemented a frame-level bit allocation algorithm for H.264 video encoding based on the Rotation Search. The obtained video quality is equivalent to other nearly-optimal algorithms, but the execution speed may be orders of magnitude faster. We also improved the unsymmetric-cross multi-hexagon-grid search (UMH) algorithm speedwise by subsampling the blocks and applying the successive elimination algorithm. The resulting method is significantly faster, requiring about 95% less time, but gives nearly the same motion estimation quality as the best algorithms.

**Machine Vision for Sensing and Understanding Human Actions**

We have continued our research on vision based human-computer interaction. Such technologies are likely to be building blocks for cognitive systems embedded in homes, offices, vehicles, and the equipment we use for everyday tasks.

A key focus has been on mobile interaction, where we have developed a method for recognizing characters from hand motions using a mobile phone. The method is based on analysis of motion trajectories, and it uses simple glyphs for representing the letters and digits. Another application investigated is controlling a mobile phone with motion patterns. A new method has been proposed whereby motion features are modelled using Hidden Markov Models (HMMs). In the experiments, eleven motion patterns were reliably recognized with this technique.

Research on using hand gesture recognition in human-computer interaction was continued. It produced a real-time demonstration, where hand gestures replaced the functions of a computer mouse by moving the cursor according to the hand trajectory and triggering the mouse functions by the recognized gestures. This method was also adopted as an interface for controlling a swarm of robots in a demonstration displayed at the Information Society Technology 2006 conference in Helsinki.

The collaboration on human action recognition with Prof. Larry S. Davis was continued with a research exchange visit by a PhD student from the University of Maryland. The joint research on ballistic human movements has produced a method that is capable of modeling and recognizing natural human movements, such as reaching and striking, which have highly variable target locations and thus huge variation in the actor’s body configuration. The method was now
extended from 3D motion capture data to real video data, and is capable of recognizing reach and strike movements from a single camera view.

Another approach to human action recognition was based on pose recognition from silhouette images using statistical models, Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). In order to use statistical models, large amounts of labeled training data are required, and to provide this, artificial data based on 3D renderings of motion capture data was used. Body parts were modeled using GMMs and then the body pose was modeled as a sequence of body parts using an HMM. Work was also done on improving the robustness of shape context features to noisy silhouette data. Good results were achieved in the recognition of five poses, training on artificial data and testing on real data.

Distance education is an application area where human action recognition is needed for selecting the appropriate video source transmitted to the remote site. In this research, we have developed an active tracking system for pan-tilt-zoom (PTZ) cameras that can track the teacher when she or he walks in front of the classroom. This system can operate robustly under large scale changes and different lighting conditions by updating color features in parallel with color-based tracking. Furthermore, it does not require any manual control, and it is able to recover from erroneous situations.

A successful combination of machine learning and bioinformatics, started in 2004, has continued. Research followed three directions: 1) feature selection based bidirectional hill-climbing (in cooperation with scientists from Portugal and Russia); 2) feature extraction using nonnegative matrix factorization (collaboration with Estonia); 3) ensembles of nearest neighbor classifiers (collaboration with Estonia). Practical tasks considered were protein fold recognition used for prediction of protein structure and gene expression based cancer classification. Dimensionality reduction, via feature extraction or feature selection, facilitated the solution of the above mentioned tasks and yielded promising results. Additionally, in the case of cancer classification, feature selection helped to focus biological verification on few selected genes instead of hundreds and thousands. Analysis of biomedical literature confirmed and complemented our findings. Ensembles (or collections) of classifiers coupled with feature selection demonstrated superior performance to a single best classifier in cancer classification. The k-nearest neighbor classifier has been studied because it is known that classifiers of this type are not easy to combine into an ensemble using traditional means. Our methodology, supported through extensive experiments with real-world data, provides the solution to this problem when a data dimensionality far exceeds the number of samples (a common setting in bioinformatics tasks).
Vision Systems Engineering

To enable useful real-world systems, our vision system engineering research provides guidelines for methodological research, helping to identify attractive approaches, architectures, and algorithms, as general purpose computing is seldom a realistic option. In practice, solutions from low-level image processing to even equipment installation and operating procedures need to be considered simultaneously. The roots of this expertise are in our industrial visual inspection research in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. Recently, we have applied our expertise to applications intended for mobile platforms, and our collaborative vision computing architecture research is a recent spin-off.

An example of a novel application intended for mobile platforms is the panorama imager that “glues” together frames selected from video sequences. The selection process analyzes for the displacements between the video frames, measures the blur due to motion and focusing, and detects moving targets ensuring that they are not mutilated in the process. In other words, the apparently simple panorama capturing process contains lots of image analysis functionality to achieve good image quality.

Machine vision applications are characterized by both high data input rates and high computational costs. For instance, a typical raw digital video rate is around 10 Mpixels/s and its processing demands at least a few hundred operations per pixel, requiring multiple GOPS of computing power. Often this needs to be done in a small package, such as a mobile communications device that may allocate at most 500 mW for application processing like video coding or person identification. This prevents implementations based on conventional processors even in the future, but hardware acceleration is needed. Currently, only monolithic long latency accelerators improve energy efficiency, but they are rigid and costly to design and verify.

Together with the Technical Universities of Tampere (Prof. Jarmo Takala) and Helsinki (Prof. Petri Vuorimaa), and the Åbo Akademi University (Prof. Johan Lilius) we are concentrating on improving the energy-efficiency of embedded high performance computing. The desired system solutions exhibit fine grained adaptable hardware accelerators to improve the silicon area efficiency via run-time reuse, while simultaneous multithreading is favored in hardware/software interfacing to reduce software overheads instead of an interrupt mechanism. We are essentially targeting on a new paradigm for embedded computing. The first step in the research has been the redesign of an MPEG-4 decoder that showed essential energy efficiency gains. A MIMO receiver solution that employs similar ideas is being developed in cooperation with the MITSE project of the Center for Wireless Communications (CWC).

In the summer Prof. Shuvra Bhattacharyya from the University of Maryland, who is the foremost expert in DSP system realizations, visited our group and provided his insight into the problems at hand. Mr. Jani Boutellier stayed with Prof. Bhattacharyya during the Fall. Both visits were very useful and enabled major steps in research. The cooperation will continue during the spring of 2007 with the visit by Mr. Tuukka Toivonen to Prof. Bhattacharyya’s group with the purpose of finding novel solutions for implementing Number Theoretic Transform (NTT) algorithms using Transport Triggered Architectures (TTA) with Synchronous Data Flow based scheduling. The NTT techniques enable exploiting correlation and convolution theorems to cut the number of memory accesses, when compared to spatial domain computations.

Visual inspection is economically still the most important application area of machine vision. The inspection systems can be relatively expensive as long as they provide high added value, and are therefore attractive proving grounds for new technologies. Typical inspection targets include part assemblies in the electronics and car industry, continuous webs such as paper, steel and fabrics, and natural materials such as wooden boards and coffee beans. Many of these targets are textured and colored, such as wood, and the inspection problem is solved best with respective methods.

Several inspection systems based on our results are marketed by our industrial partners for applications ranging from coffee bean sorting to paper formation measurement. Recent systems employ both our visual training approach and LBP based texture analysis methods. Probably the most notable implementation of our technologies is the Surface Quality Monitoring tool marketed by Cognex Inc. The actual development work has been carried out by Intopii Oy, Oulu.

Currently, we are investigating methods and means for building visual inspection systems for exceptionally demanding applications such as non-destructive, non-contact dynamic strength grading of wooden boards. The underlying observation is that the strength is different to the axial, tangential and radial directions of the wood grain, while the behavior of the grain is affected differently by sound and dry knots. After color and texture based image analysis has provided the grain and knot information, a 3-D Finite Element Model (FEM) is built and analyzed for strength. Due to the interdisciplinary nature of the research area, the MVG has started cooperation discussions with Prof. Mark Hughes’ group at the Wood Technology Laboratory of the Helsinki University of Technology.

The dynamic ranges of the camera sensors are rather limited. As a cure, we have endeavored to create techniques that enable the capturing of good quality images from unevenly illuminated scenes. With a high frame rate camera this could be done, for example, using alternating integration times for sequential frames. The proper integration times could be determined via content analysis, for example, by evaluating the quality of human “skin signal”. Our good
preliminary results suggest that this kind of image acquisition technology could be useful in proactive systems and visual surveillance applications.

**Exploitation of Results**

Our approach of combining world-class basic research with more applied research on vision systems and systems engineering is quite unique, giving rise to our research having a great practical impact. We conceive of machine vision research as a remarkable field of science that contributes to the competitiveness of Finnish enterprises by developing methods and techniques for improving the performance and usability of industrial machines and products.

In its April 2006 issue, American Vision Systems Design magazine highlighted the research on wood inspection technology led by Prof. Silvén, which has been commercialized by inX Systems as an Opti-Grader board quality measurer and used by UPM in its mills.

In addition, the results of the project on developing compiler technology of the embedded systems are expected to have a significant commercial influence. The new compiler technology provides significantly better energy efficiency for the embedded systems. It also provides new ways to develop systems when the compiler is able to take some responsibilities from the layered architecture, operating system, and the application designer. New technology also puts an end to the ever-growing complication of the layered architectures.

We are working together with Videra Oy to enable home video sensor technologies for detecting accidents and illness related problems. The envisioned systems combine learning technologies with image analysis, and are intended to be easily installable in almost any home. The first uses of the technology are expected to be in retirement and nursing homes.

Another example of the impact of our work is that in 2005 Intopii Oy, a spin-off company of our texture research, entered into a cooperative agreement with the Cognex Corporation, the world’s leading supplier of machine vision systems. Late 2006 saw the launch of another spin-off, Visidon Oy, that intends to specialize on applications running on mobile platforms.

**Future Goals**


In order to develop challenging real-world applications, many scientific and engineering problems need to be solved. Although very successful in controlled environments, as in the industry, to reach homes and consumers, machine vision needs major breakthroughs and generic methodologies that make the technology inherently robust and simple to use. In practice, the Machine Vision Group will be able to achieve this by keeping the successful approach of carrying out top-class fundamental research in chosen key areas, close interaction between basic and applied research, and in-depth international collaboration. Our more active participation in European research programs can be expected within the 7th framework research program of the EU.

**Personnel**

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>12</td>
</tr>
<tr>
<td>graduate students</td>
<td>17</td>
</tr>
<tr>
<td>others</td>
<td>12</td>
</tr>
<tr>
<td>total</td>
<td>41</td>
</tr>
<tr>
<td>person years</td>
<td>27</td>
</tr>
</tbody>
</table>

**External Funding**

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>241 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>218 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>268 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>99 000</td>
</tr>
<tr>
<td>total</td>
<td>826 000</td>
</tr>
</tbody>
</table>

**Doctoral Theses**

Selected Publications


Niskanen M (2006) View dependent enhancement of the dynamic range of video. Proc. 18th International Conference on Pattern Recognition (ICPR 2006), Hong Kong, 1: 984-987


Background and Mission

The Machine Vision Group (MVG) is renowned world-wide for its expertise in computer vision. It works as a single well-focused research group and collaborates with other groups of similar status in Europe, the USA and China. The research areas of the group range from generic computer vision methodologies to machine vision applications and vision systems engineering. The results of its research have been widely exploited in industry. The current areas of interest include face analysis, camera-based user interfaces, visual surveillance, smart environments and energy-efficient architectures for vision computing.

Highlights and Events in 2007

During the year 2007, MVG participated in two thorough research evaluations with international evaluator committees, organized by the Academy of Finland and the University of Oulu. The results of both of the evaluations show that the chosen research strategies and practices have been successful, and that the research team has established its position among the most significant research teams in the field of machine vision.

Two important events were organized. First, in June 2007, the group organized the 30th anniversary seminar of the Pattern Recognition Society of Finland. The keynote speaker of the event was Prof. Rama Chellappa from the University of Maryland, who is one of the leading researchers in the field.

In August 2007, the group organized the Finnish Signal Processing Symposium. Participants of the event included mainly graduate students from Finland, but also some students from abroad took part in the symposium. The invited speech of the event was given by Prof. Yrjö Neuvo.

The group has attracted public visibility both within and outside academic forums. During the year 2007, the group made its new record in the amount of scientific publications. Also the quality of the publications has traditionally been good, and the papers have been published in leading conferences and journals of the field.

In addition, visibility outside the academic forums has been reached. The group has presented its research projects and activities several times in the most significant local and national, and also some international media forums presenting science news in the field. Among these forums were, for example, a national television programme Prisma Studio, the regional news of YLE, the newspaper Kaleva and a national professional magazine Prosessori.

The group has hosted visits of several respected and well-known scientists from abroad, including Professor Thomas S. Huang from the University of Illinois at Urbana-Champaign, Professor Rama Chellappa from the University of Maryland, and a group of researchers from the Academy of Sciences of the Czech Republic led by Professor Jan Flusser. In addition, several domestic visitors, for example from the Academy of Finland, were briefed about our research activities.

The group has established active collaboration with some of the world’s leading institutions and top scientists. It has had in-depth collaboration with the University of Maryland (USA) since the early 1980’s, and more recent partners include the Chinese Academy of Sciences, INRIA Rhone-Alpes (France), and the Academy of Sciences of the Czech Republic. Joint research efforts with the University of Freiburg (Germany) have also been made.

The group fosters researcher mobility to and from our unit. Two of our researchers made research visits to partner institutions during the reporting period. The group has also visiting postdoctoral researchers and graduate students from abroad. Recently the co-operation with Chinese universities has been intensifying, and during the year 2007 two
new Chinese postdoctoral researchers joined the group. This mobility has led to joint research projects and co-authored publications.

The group and its members are active in the scientific community. For example, in 2006-2007 Prof. Pietikäinen served as an area chair of the top-ranking IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007). He is also a Co-Chair of Workshops for the International Conference on Pattern Recognition (ICPR 2008). The professors of the group were committee members of several other international conferences and many researchers of the group served as reviewers for various journal and conference articles.

Scientific Progress

The current focus areas of the research consist of: 1) texture-based computer vision, including facial image analysis 2) geometric image and video analysis 3) machine vision for sensing and understanding human actions, 4) learning in machine vision, and 5) vision systems engineering, including new paradigms for embedded systems.

Texture-Based Computer Vision

Texture is a fundamental property of surfaces. It can be seen almost anywhere. We have developed a novel methodology based on Local Binary Patterns (LBP), which has evolved to present a major breakthrough in texture analysis. It is already widely used all over the world. Our recent results show that the approach can offer significant potential for many important tasks in computer vision which have not been earlier regarded as texture problems.

In 2007, our research focused on robust LBP-based descriptors for static and dynamic textures, and on texture-based methods for face detection, face recognition, facial expression recognition, visual speech recognition and recognition of actions.

In a study aiming for a better understanding of the properties of the Local Binary Pattern operator, a framework for image descriptors based on quantized joint distribution of filter bank responses was formulated. In this framework, it was shown that despite their small spatial support, the oriented derivative filters that can be used to implement the LBP operator outperform Gabor and MR8 filters in the texture categorization task. Furthermore, codebook and thresholding based quantization of filter bank responses were compared. Even though conceptually and computationally simpler, thresholding was found to perform better than codebook based quantization in many cases.

Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. Earlier, a novel Center-Symmetric-LBP (CS-LBP) interest region descriptor which combines the strengths of the SIFT descriptor and the LBP texture operator was proposed. In extensive experiments, our descriptor performed better than SIFT for most of the test cases, and about equally well for the remaining ones. Especially the tolerance of our descriptor to illumination changes was clearly demonstrated. Lately, we have prepared and submitted a journal article on our method. This research was carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhône-Alpes, France.

The work with texture analysis in industrial applications was continued. An enhanced method for training SOM maps for defect detection with the LBP texture feature was developed. A journal article on this topic is in revision.

In 2005, we began to study methods for analyzing dynamic textures, i.e. textures in motion. A volume LBP operator (VLBP) was developed which combines temporal and spatial information (i.e. motion and appearance). A simplified method based on concatenated LBP histograms computed from three orthogonal planes (LBP-TOP) was also proposed. In 2007, a paper on these descriptors with an application to facial expressions was published in the prestigious IEEE Transactions on Pattern Analysis and Machine Intelligence journal. In the reporting year, we studied some improvements of the original VLBP operator and presented a proof on the relation of two versions of rotation invariant VLBP patterns.

In a study aiming for a better understanding of the properties of the Local Binary Pattern operator, a framework for image descriptors based on quantized joint distribution of filter bank responses was formulated. In this framework, it was shown that despite their small spatial support, the oriented derivative filters that can be used to implement the LBP operator outperform Gabor and MR8 filters in the texture categorization task. Furthermore, codebook and thresholding based quantization of filter bank responses were compared. Even though conceptually and computationally simpler, thresholding was found to perform better than codebook based quantization in many cases.

Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. Earlier, a novel Center-Symmetric-LBP (CS-LBP) interest region descriptor which combines the strengths of the SIFT descriptor and the LBP texture operator was proposed. In extensive experiments, our descriptor performed better than SIFT for most of the test cases, and about equally well for the remaining ones. Especially the tolerance of our descriptor to illumination changes was clearly demonstrated. Lately, we have prepared and submitted a journal article on our method. This research was carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhône-Alpes, France.

The work with texture analysis in industrial applications was continued. An enhanced method for training SOM maps for defect detection with the LBP texture feature was developed. A journal article on this topic is in revision.

In 2005, we began to study methods for analyzing dynamic textures, i.e. textures in motion. A volume LBP operator (VLBP) was developed which combines temporal and spatial information (i.e. motion and appearance). A simplified method based on concatenated LBP histograms computed from three orthogonal planes (LBP-TOP) was also proposed. In 2007, a paper on these descriptors with an application to facial expressions was published in the prestigious IEEE Transactions on Pattern Analysis and Machine Intelligence journal. In the reporting year, we studied some improvements of the original VLBP operator and presented a proof on the relation of two versions of rotation invariant VLBP patterns.

In a study aiming for a better understanding of the properties of the Local Binary Pattern operator, a framework for image descriptors based on quantized joint distribution of filter bank responses was formulated. In this framework, it was shown that despite their small spatial support, the oriented derivative filters that can be used to implement the LBP operator outperform Gabor and MR8 filters in the texture categorization task. Furthermore, codebook and thresholding based quantization of filter bank responses were compared. Even though conceptually and computationally simpler, thresholding was found to perform better than codebook based quantization in many cases.

Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. Earlier, a novel Center-Symmetric-LBP (CS-LBP) interest region descriptor which combines the strengths of the SIFT descriptor and the LBP texture operator was proposed. In extensive experiments, our descriptor performed better than SIFT for most of the test cases, and about equally well for the remaining ones. Especially the tolerance of our descriptor to illumination changes was clearly demonstrated. Lately, we have prepared and submitted a journal article on our method. This research was carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhône-Alpes, France.

The work with texture analysis in industrial applications was continued. An enhanced method for training SOM maps for defect detection with the LBP texture feature was developed. A journal article on this topic is in revision.

In 2005, we began to study methods for analyzing dynamic textures, i.e. textures in motion. A volume LBP operator (VLBP) was developed which combines temporal and spatial information (i.e. motion and appearance). A simplified method based on concatenated LBP histograms computed from three orthogonal planes (LBP-TOP) was also proposed. In 2007, a paper on these descriptors with an application to facial expressions was published in the prestigious IEEE Transactions on Pattern Analysis and Machine Intelligence journal. In the reporting year, we studied some improvements of the original VLBP operator and presented a proof on the relation of two versions of rotation invariant VLBP patterns.

In a study aiming for a better understanding of the properties of the Local Binary Pattern operator, a framework for image descriptors based on quantized joint distribution of filter bank responses was formulated. In this framework, it was shown that despite their small spatial support, the oriented derivative filters that can be used to implement the LBP operator outperform Gabor and MR8 filters in the texture categorization task. Furthermore, codebook and thresholding based quantization of filter bank responses were compared. Even though conceptually and computationally simpler, thresholding was found to perform better than codebook based quantization in many cases.

Local photometric descriptors computed for regions around interest points have proven to be very successful in many problems. These local features are distinctive, and robust with respect to changes in viewpoint, scale and occlusion. The most widely used is the SIFT descriptor proposed by David Lowe. Earlier, a novel Center-Symmetric-LBP (CS-LBP) interest region descriptor which combines the strengths of the SIFT descriptor and the LBP texture operator was proposed. In extensive experiments, our descriptor performed better than SIFT for most of the test cases, and about equally well for the remaining ones. Especially the tolerance of our descriptor to illumination changes was clearly demonstrated. Lately, we have prepared and submitted a journal article on our method. This research was carried out in collaboration with Dr. Cordelia Schmid from INRIA Rhône-Alpes, France.

The work with texture analysis in industrial applications was continued. An enhanced method for training SOM maps for defect detection with the LBP texture feature was developed. A journal article on this topic is in revision.
In 2004, we proposed a novel facial representation based on LBP features, obtaining excellent results. A paper on this topic was published in 2006 in IEEE Transactions on Pattern Analysis and Machine Intelligence. Our approach has evolved to be a growing success. It has been adopted and further developed by many research groups and leading scientists working in the field.

We continued the investigation of spatiotemporal approaches to face recognition from videos by considering an extended set of Volume LBP features and AdaBoost learning. Among the key properties of the new approach are: (i) the use of local Extended Volume LBP based spatiotemporal description instead of the holistic representations commonly used in previous works; (ii) the selection of only personal specific facial dynamics while discarding the intrapersonal temporal information; and (iii) the incorporation of the contribution of each local spatiotemporal information. The extensive experimental analysis with three different publicly available video face databases and five benchmark methods (PCA, LDA, LBP, HMMs and ARMA) clearly assessed the excellent performance of the proposed approach, significantly outperforming the comparative methods and thus advancing the state-of-the-art. We also began the application of the proposed approach to other facial image analysis tasks such as gender classification from videos.

Research and experiments on face and eye detection for person authentication in mobile phones were also made yielding in a prototype of an authentication system using Haar-like features with AdaBoost for face and eye detection, and LBP features with histogram intersection matching for face verification. Average authentication rates of 82% for small-sized faces (40×40 pixels) and 96% for faces of 80×80 pixels were obtained on a Nokia N90 mobile phone with an ARM9 processor with 220 MHz. These obtained results are encouraging, and point to the feasibility of face authentication in mobile phones.

Research on face recognition techniques was also continued with the development of a pattern recognition database that would make it possible to classify faces against large databases in real time. Extensive tests with a large-scale similarity matching study were made. An article was written for a journal (yet to be published).

Facial expression recognition using LBP-TOP descriptors was further studied. It was shown that our approach performs well also at low frame rates and for low-resolution image sequences. A near real-time experimental system was implemented to demonstrate the applicability of the method. A simple web camera was used to capture videos in an office environment.

We also continued research on recognition of isolated phrases using only visual information. Visual speech information plays an important role in speech recognition under noisy conditions or for listeners with hearing impairment. A human listener can use visual cues, such as lip and tongue movements, to improve speech understanding. An appearance feature representation based on LBP-TOP descriptors was proposed, taking into account the motion of the mouth region and time order in pronunciation. Positions of the eyes determined by a robust face and eye detector are used for localizing the mouth regions in face images. Spatiotemporal LBP-TOP patterns extracted from these regions are applied for describing phrase sequences. In our experiments, promising accuracies of 62% and 70% were obtained in speaker-independent and speaker-dependent recognition, respectively.

In comparison with other methods on the Tulips1 audio-visual database, the accuracy of 92.7% of our method clearly outperformed the others. Advantages of our approach include local processing and robustness to monotonic grayscale changes. Moreover, no error prone segmentation of moving lips is needed.

Feature definition and selection are two important aspects in visual analysis of motion. We have also been investigating the use of the spatiotemporal local binary patterns computed at multiple resolutions for describing dynamic events, which combine static and dynamic information from different spatiotemporal resolutions. Appearance and motion are the key components for visual analysis related to movements. The AdaBoost algorithm is utilized for learning the principal appearance and motion from spatiotemporal descriptors derived from three orthogonal planes, providing important information about the locations and types of features for further analysis. In addition, learners are designed for selecting the most important features for each specific pair of different classes. The figure above shows the selected features for two expression pairs. The experiments carried out on diverse visual analysis tasks, facial expression recognition and visual speech recognition, show the effectiveness of the approach.

**Geometric Image and Video Analysis**

Imaging geometry provides the basic mathematical tools for analyzing the spatial properties of images. For image sequences, motion is another important feature that describes the temporal changes between successive images. Topics dealing with both imaging geometry and motion have been explored in this part of the research.

During the year 2007, we have continued our research in geometric camera calibration. The emphasis has been on the modeling and calibration of omnidirectional cameras which is currently an active research area in the computer vision community. In particular, we have further developed our generic calibration approach, which is based on using a planar calibration pattern and a flexible parametric camera model. The approach was experimented on with various kinds of cameras, including fish-eye lens cameras and catadioptric cameras, and it was observed that in most cases a subpixel level of calibration accuracy could be achieved.
The results of the development work were included in our camera calibration toolbox which is publicly available on the Internet, and already widely used in different universities and organizations all over the world. Furthermore, our first studies on self-calibration of generic omnidirectional cameras have been published. Potential application areas of this research include panoramic imaging and 3D modeling.

Recent studies and advances in image-based 3D reconstruction and object recognition suggest that the reconstruction and recognition procedures should be combined in order to build efficient systems for automatic scene analysis. Inspired by this development we have extended our previous work about quasi-dense matching, and applied it for object recognition tasks. The results obtained with publicly available datasets indicate that reliable object recognition is possible, also in difficult viewing conditions where extensive background clutter, partial occlusion, large scale and viewpoint changes, or notable geometric deformations are present.

Our research on affine invariants concentrated on finalizing the formulation of the multiscale framework and expanding the variety of the related image transformations. The methodology of the presented framework is provided by the idea of combining novel image transformation to the existing feature extractors in both recognition and registration, which enables us to effectively multiply the number of output features of the original methods. A major advantage in the approach is also the possibility to use one transformation in combination with many different feature extractors, enabling a generic use of the developed techniques. The results presented in the affine case hold great promise for generalization to cover other distortions as well.

In many practical situations, where imaging conditions are not ideal, the images captured are degraded by blur due to motion of the camera or the lens system being out of focus. This problem has been addressed in our research on blur invariant pattern recognition and image registration. The frequency domain invariant features developed in our group are insensitive to centrally symmetric blur, including linear motion and out of focus blur. In our recent work, the blur invariant phase correlation (BIPC) method has been extended to images that are also subject to rotation, scaling and translation. Similar methodology has been used for object recognition purposes as well. Our latest achievement was to broaden the method for affine invariance. In all these cases, our methods have outperformed the blur moment invariants that are the only previously known approach to deal with image blurring in object recognition and image registration.

Another problem related to common acquisition devices is that they are not able to capture the entire dynamic range that natural scenes often exhibit. High Dynamic Range Imaging (HDRI) techniques try to recover the radiance at each pixel location by merging a sequence of images taken with different exposure times. However, results are often spoiled by the presence of moving objects that cause ghosting artifacts in the output image. A new ghost-removal approach has been developed in order to make HDR acquisition of non-static scenes more feasible. The proposed method proved to be fairly robust and work better than existing algorithms in many circumstances.

In our research on image processing for an MRI-compatible robot, usage of computer vision for surgical navigation has been studied, and an experimental setup with near-IR lighting and cameras has been built. Different camera calibration and pose estimation methods and algorithms have been implemented and evaluated. Based on this research, we have developed a marker based pose estimation framework, which can be utilized in the control of a surgical robot.

**Machine Vision for Sensing and Understanding Human Actions**

We have continued our research on vision based human-computer interaction. Such technologies are likely to be building blocks for cognitive systems embedded in homes, offices, vehicles, and the equipment we use for everyday tasks.

Our research on vision-based human-computer interaction has progressed with several new advanced concepts. One of them is a technique for finger tracking with a mobile device, where we utilize a novel method that embeds the Kalman filter and the expectation maximization algorithms.
in a special way to estimate the finger as well as the background motion. This Kalman-EM technique was successfully applied for controlling the user interface of a mobile phone, where the users were allowed to interact with the device by moving a finger in front of the camera.

Interaction with a mobile phone by moving a finger in front of the camera.

Finger motion can be directly used to browse information on the display of the device, but it also enables us to give specific commands to the device with simple gestures, which involves using hidden Markov models (HMMs) to model sequences of motion features. In order to improve the recognition performance we have proposed an approach, where the motion trajectories are filtered based on the entropy of a histogram of the velocity. Sequences with high entropy, and so more random velocity, are rejected as possibly being unintended or incorrect. This work was extended to adapting the models to individual users using unsupervised maximum a posteriori (MAP) adaptation. The velocity entropy measure was used as criteria for successfully controlling the adaptation process.

Another related topic is the research on a camera based document scanner that has been conducted in collaboration with the University of Maryland. Instead of using devices such as flatbed scanners, our solution was to allow the users to capture large document images with their mobile phones equipped with a camera. In this work, we have developed a method where the device interactively guides the user to move the device over, for example, a newspaper page in a manner that a panorama image can be assembled from individual frames. During online scanning, motion determined from low-resolution images is used to control the interaction process, while good high-resolution images of the document page are captured from the most favorable locations and used to construct the mosaic image automatically.

Automatic interpretation of hand gestures has many potential applications, for example in natural user interfaces, automatic sign language recognition, virtual reality, and even emotion recognition. We developed a robust real-time method for hand tracking based on particle filtering. The method uses computationally efficient color blob features for effective propagation of particles. The experiments showed that the method is able to track a hand in the presence of complicating factors such as fast hand movement, and clutter and movements in the background.

Hand gestures can act as commands in natural user interfaces.

Our work has continued on the recognition of human body parts from silhouette images based on statistical models, specifically, Gaussian mixture models and hidden Markov models. Using large amounts of artificially generated labeled training data, models were created for human body parts. The performance of these models was tested using a variety of real test data collected from different sources. These tests showed that the models performed well with very noisy and occluded silhouettes. This body part recognition was extended to unusual pose recognition by estimating the overall model confidence based on the likelihood ratio of the recognized pixels.

We have also developed a new algorithm for tracking multiple objects. The method is based on our Kalman-EM framework, where the measurements are assumed to obey a dynamic Gaussian mixture model. The algorithm includes a novel way of extracting the measurements from binary masks using basic morphological operations. In the current implementation, we utilize color features to determine the interesting objects from the background. Preliminary experiments are promising and they indicate that the algorithm could have great potential in practical multi-object tracking applications.

Human motion can be seen as a type of moving texture pattern. We developed a novel description for human movements by describing human activities with texture features. We use temporal templates as a pre-processing stage and describe their local characteristics with LBP features to obtain a short time motion description. By using the local properties of movements, our method captures the essential information from human movements while it also allows variation in performance of activities. The method was tested on a database (http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html) of various human movements. A classification rate of 98% was achieved with HMM modeling, which equals the best results reported.

A Tekes funded project PersonID came to an end in the summer of 2007. During the last year of the project, a state-of-the-art image enhancement method based on blind super-resolution was applied to real world video surveillance data. Also, the surveillance algorithm developed in the previous year of the same project was developed further: Ab-
normal human activity detection was studied and, as a result, an algorithm capable of detecting simple but important anomalies in human behavior (like collapsing and tripping) was added to the existing multi-object tracking framework. After the Tekes project ended, the research focus has been on object tracking and recognition in a distributed environment of multiple cameras and image processing elements.

Learning in Machine Vision

Learning is a key bottleneck in building cognitive machine vision systems, regardless of whether they are intended for industrial inspection, visual surveillance, or medical image analysis. Most of our recent research on learning is related to manifold learning and nonlinear dimensionality reduction, as it is known that high-dimensional feature data often lies in a lower dimensional manifold in the observation space. This can be exploited by learning the low dimensional manifolds from image data, and then selecting appearance models or features for recognition.

This approach is used as such in interdisciplinary brain activity analysis research which utilizes functional magnetic resonance imaging (fMRI). This work was initiated in 2006 together with the Department of Radiology (the largest unit in the Nordic countries) at the University of Oulu, and the Chinese Academy of Sciences (Prof. Yu-Feng Zang, National Laboratory of Pattern Recognition). Our goal is to develop clinically applicable diagnostic methods modeled according to solutions originally developed for learning in visual inspection applications. The goal is the early detection of neurodegenerative diseases such as Parkinson’s and Alzheimer's. Resting state fMRI (functional Magnetic Resonance Imaging) is used for obtaining the data. A project on this topic is funded from the NEURO 2006-2009 program of the Academy of Finland.

A successful combination of machine learning and bioinformatics, started in 2004, has continued. A new non-negative matrix factorization algorithm for unsupervised data reduction was proposed and tested on face images. Its goal is to detect most representative patterns for each class of data without knowing a priori class labels of patterns. These representative patterns can facilitate clustering or classification of large data sets. Results were published in the Signal Processing Journal.

Research on ensembles (collections) of classifiers led to the 1st Workshop on Supervised and Unsupervised Ensemble Methods and their Applications organized by Dr. Oleg Okun together with Dr. Giorgio Valentini from the University of Milan, Italy, in Girona, Spain on June 4, 2007. Both co-chairs acted as editors of the workshop proceedings. The workshop was attended by researchers from Spain, Portugal, Italy, France, the USA, and Finland.

Two classifier ensembles (of k-nearest neighbors and of decision trees) were researched for gene expression based cancer classification, with the conclusion that ensembles of k-nearest neighbors are much smaller in size and more accurate than those of decision trees.

Vision Systems Engineering

To enable useful real-world systems, our vision system engineering research provides guidelines for methodological research, helping to identify attractive approaches, architectures, and algorithms, as general purpose computing is seldom a realistic option. In practice, solutions from low-level image processing to even equipment installation and operating procedures need to be considered simultaneously. The roots of this expertise are in our industrial visual inspection research in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. Recently, we have applied our expertise to applications intended for mobile platforms, and our collaborative vision computing architecture research is a recent spin-off.

An example of a novel application intended for mobile platforms is the panorama imager that “glues” together frames selected from video sequences. The selection process analyzes the displacements between the video frames, measures the blur due to motion and focusing, selecting the suitable frames for mosaicing, and detects moving targets and human faces ensuring that they are not mutilated in the process. In other words, the apparently simple panorama capturing process contains lots of image analysis functionality to achieve good image quality. Some of the solutions developed for mobile platforms are already being re-used in industrial applications. For instance, the frame selection techniques of panorama capture are employed in developing a matrix camera based quality monitoring system for a printing machine that can cope with flutter and frequent environmental disturbances.

Machine vision applications are characterized by both high data input rates and high computational costs. For instance, a typical raw digital video rate is around 10 Mpixels/s and its processing demands at least a few hundred operations per pixel, requiring multiple GOPS of computing power. Often this needs to be done in a small package, such as a mobile communications device that may allocate at most 500 mW for application processing like video coding or person identification. This prevents implementations based on conventional processors even in the future, but hardware acceleration is mandatory. Currently, only monolithic long latency accelerators improve energy efficiency, but they are rigid and costly to design and verify, and difficult to justify for purposes which are considered marginal and computationally expensive.
Together with the Technical Universities of Tampere (Prof. Jarmo Takala) and Helsinki (Prof. Petri Vuorimaa), and the Åbo Akademi University (Prof. Johan Lilius) we are concentrating on improving the energy-efficiency of embedded high performance computing. We have demonstrated that fine grained, silicon area efficient adaptable hardware accelerators can be employed at very low software interface overheads through deterministic multithreaded schedules. This has turned out to be much more efficient than the conventional interrupt, semaphore, and event handler mechanisms advocated by the textbooks. In essence, we are targeting a new paradigm for embedded computing and expect significant impacts in the field. Our current demonstrations include the simultaneous decoding of multiple MPEG-4 streams on shared accelerators and MIMO reception. The latter work has been carried out in cooperation with Centre for Wireless Communications (CWC).

Visual inspection is economically still the most important application area of machine vision. The inspection systems can be relatively expensive as long as they provide high added value, and are therefore attractive testing grounds for new technologies. Typical inspection targets include part assemblies in the electronics and car industry, continuous webs such as paper, steel and fabrics, and natural materials such as wooden boards and coffee beans. Many of these targets are textured and colored, such as wood, and the inspection problem is solved best with respective methods. Several inspection systems based on our results are marketed by our industrial partners for applications ranging from coffee bean sorting to paper formation measurement.

Currently, we are investigating methods and means for building visual inspection systems for exceptionally demanding applications such as non-destructive, non-contact dynamic strength grading of wooden boards. The underlying observation is that the strength is different to the axial, tangential and radial directions of the wood grain, while the behavior of the grain is affected differently by sound and dry knots. After color and texture based image analysis has provided the grain and knot information, 3-D Finite Element Model (FEM) is built and analyzed for strength. Due to the interdisciplinary nature of the research area, the MVG has started cooperation discussions with Prof. Mark Hughes’ group at the Wood Technology Laboratory of the Helsinki University of Technology.

The dynamic ranges of the camera sensors are rather limited. As a cure, we have endeavored to create techniques that enable the capturing of good quality images from unevenly illuminated scenes, and under infra-red illumination. These techniques borrow from our visual inspection and mobile device imaging solutions, once more bridging between the apparently very different domains.

**Exploitation of Results**

Our approach of combining world-class basic research with more applied research on vision systems and systems engineering is quite unique, giving rise to our research having a great practical impact. We conceive of machine vision research as a remarkable field of science that contributes to the competitiveness of Finnish enterprises by developing methods and techniques for improving the performance and usability of industrial machines and products.

The results of the project on developing novel solutions for embedded systems design are expected to have a significant commercial influence. The new technology provides significantly improved energy efficiency when fine grained hardware acceleration is mandatory. The first commercial uses are expected to be in mobile video codecs. We are...
working together with Videra Oy to enable home video sensor technologies for detecting accidents and illness related problems. The solutions synthesize infrared imaging, human action recognition, and visual learning technologies that make the systems installable in almost any home. The first uses of the technology are expected to be in retirement and nursing homes.

Another example of the impact of our work is that in 2005 Intopi Ltd., a spin-off company from our texture research, entered into a cooperative agreement with the Cognex Corporation, the world’s leading supplier of machine vision systems. In late 2006 another spin-off, Visidon Ltd., was launched. The company provides intelligent computer vision solutions for mobile devices as well as special system, algorithm, and software design and training services for a variety of demanding industrial and consumer applications.

Future Goals

The Machine Vision Group is now stronger than ever. Working as a single well-focused research group, in which different teams and researchers work closely together, has made it possible for them to benefit from each other’s work and cumulative past experiences in an efficient way.

Within international collaboration, the group will participate in a new European FP7 project called Mobile Biometry (MOBIO) coordinated by the IDIAP Research Institute, Switzerland. The goal of this three-year project (2008-2010) is to investigate multiple aspects of face and speech data in user authentication of mobile devices. The approach for face description using local binary patterns introduced by MVG, will play a significant role in the consortium.

The research activities of the group have always been a proven ground on actual research challenges and improving the state-of-art methods. In the near future, more attention will be given to ubiquitous computing, where machine vision applications are being embedded into the home environment even as invisible solutions.

It is becoming technically possible to build smart rooms which have an all-around technical wireless infrastructure capable of sensing and interpreting human actions. Machine vision will play a key role in developing such ubicomp systems. Our successful research on sensing and understanding human actions, face detection and recognition, and interpretation of emotions, as well as energy efficient engineering continues within new European, Academy and Tekes projects.

In order to develop challenging real-world applications, many scientific and engineering problems need to be solved. Although very successful in controlled environments in the industry, to reach homes and consumers, machine vision needs major breakthroughs and generic methodologies that make the technology inherently robust and simple to use. In practice, the Machine Vision Group will be able to achieve this by preserving the successful approach of carrying out top-class fundamental research in chosen key areas, close interaction between basic and applied research, and in-depth international collaboration.

Personnel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>13</td>
</tr>
<tr>
<td>graduate students</td>
<td>15</td>
</tr>
<tr>
<td>others</td>
<td>15</td>
</tr>
<tr>
<td>total</td>
<td>43</td>
</tr>
<tr>
<td>person years</td>
<td>28</td>
</tr>
</tbody>
</table>

External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>242 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>210 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>228 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>355 000</td>
</tr>
<tr>
<td>total</td>
<td>1 035 000</td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


Background and Mission

The Machine Vision Group (MVG) is renowned world-wide for its expertise in computer vision, due, for example, to its highly successful Local Binary Pattern (LBP) methodology. The group has a solid record of scientific merits on both basic and applied research on computer vision which now spans 27 years. The mission of the group is to always pursue actual research challenges and improve the state-of-the-art methods.

MVG works as a single well-focused research group and intensively collaborates with other groups of similar status in Europe, the USA and China. The group has been invited to take part in European project proposals and has a joint research project on face analysis and visual surveillance together with Prof. Stan Z. Li from the Institute of Automation at the Chinese Academy of Sciences. These are clear signals of our attractiveness as a distinguished partner in the global research community. Within the Seventh Framework Programme FP7, the group currently participates in a project consortium of Mobile Biometry (MOBIO), which is coordinated by the IDIAP Research Institute, Switzerland.

The research areas of the group range from generic computer vision methodologies to machine vision applications and vision systems engineering. The results of our research have been widely exploited in industry, and contract research forms a part of our activities. The current areas of research interest include texture analysis, geometric image and video analysis, face analysis, analysis of motion and human actions, camera based user interfaces, visual surveillance, and energy-efficient architectures for vision computing.

Highlights and Events in 2008

During 2008, the group produced a large number of significant scientific contributions and published its key results through major conferences and journals. The group was very successful in winning highly competitive funding for its long term research from the Academy of Finland. Altogether, four new Academy projects were launched in January 2009. This shows that the research group has established its position among the most significant research teams in the field of machine vision. In addition, visibility outside the academic forums has been achieved. The group has a tradition of presenting regularly its research projects and activities in different media. Among these were this year, for example, a popular nationwide magazine Tiede, a national professional magazine Prosessori, the Northern Finland news of YLE, the newspaper Kaleva and many online releases on technology.

The group hosts regularly visits of respected and renowned scientists from abroad. In 2008, the group had the pleasure of hosting Professor Josef Kittler from the University of Surrey, Professor Jan Flusser from the Academy of Sciences of the Czech Republic, and Professor Jiri Matas from the Czech Technical University in Prague. In addition, several domestic visitors from both public and private sectors were briefed about our research activities.

The group has established active collaboration with some of the world’s leading institutions and top scientists. It has had collaboration with the University of Maryland (USA) from the early 1980’s. More recent partners include the Chinese Academy of Sciences, the Academy of Sciences of the Czech Republic, Czech Technical University in Prague, the École Polytechnique Fédérale de Lausanne (EPFL), and the consortium of the European Mobile Biometry (MOBIO) project. The eight partners of the MOBIO consortium convened in Oulu in July, 2008.

The group fosters international mobility to and from our unit. Five of our researchers made research visits to foreign institutions during the reporting period. The group has attracted visiting postdoctoral researchers and postgraduate students from abroad, who are affiliated with us for periods of a couple of weeks up to several years.

The group and its members are active in the scientific community. For example, in 2008 Prof. Pietikäinen co-chaired the workshops for the International Conference on Pattern Recognition (ICPR 2008), and he and Dr. Guoying Zhao visited Oulu in summer 2008.
served as co-chairs of the ECCV 2008 workshop on Machine Learning for Vision based Motion Analysis (MLVMA 08) together with Dr. Liang Wang and Dr. Li Cheng. Dr. Abdenour Hadid presented a tutorial on Face Analysis using Local Binary Patterns at the International Workshops on Image Processing Theory, Tools and Applications. The professors of the group were committee members of several major conferences and many researchers of the group served as reviewers for various journal and conference articles. Dr. Guoying Zhao has been serving as a co-editor of the forthcoming book titled “Machine learning for human motion analysis: theory and practice”, IGI Global, 2009, together with Dr. Liang Wang and Dr. Li Cheng.

Scientific Progress

The current main areas of the research are: 1) computer vision methods, 2) human-centered vision systems, and 3) vision systems engineering.

Computer Vision Methods

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis, and geometric image and video analysis. The basic research in these areas has created the basis for many novel contributions in our research on human-centered vision systems and vision systems engineering.

Texture Analysis

Texture is a fundamental property of surfaces. It can be seen almost anywhere. We have developed a novel methodology based on Local Binary Patterns (LBP), which has evolved to present a breakthrough in texture analysis. It is frequently cited and widely used all over the world. In 2008, our research focused on robust LBP based descriptors for static and dynamic textures, and on texture based methods for different tasks in face and activity analysis.

In a study aiming for a better understanding of the properties of the local binary pattern operator, a unified framework for image descriptors based on quantized joint distribution of filter bank responses was formulated, and the significance of the filter bank and vector quantizer selection was evaluated. A filter bank based representation of the Local Binary Pattern (LBP) operator was introduced, showing that LBP can also be presented as an operator producing vector quantized filter bank responses. Maximum Response 8 (MR8) and Gabor filters are widely used alternatives to the derivative filters which are used to implement LBP. The performance of these three sets was compared in texture categorization and face recognition tasks.

A novel Bayesian LBP operator was proposed. This operator is formulated in a new Filtering, Labeling and Statistic (FLS) framework for texture descriptors. In the framework, the local labeling procedure, which is a part of many popular descriptors such as LBP, SIFT and VZ, can be modeled as a probability and optimization process. This enables the use of more reliable prior and likelihood information, and reduces the sensitivity to noise. The BLBP operator pursues a label image, when given the filtered vector image, by maximizing the joint probability of two images under the criterion of MAP. The proposed approach was evaluated on texture retrieval schemes using the entire Brodatz database. The result reveals the BLBP operator’s efficient performance and FLS framework’s capability for in-depth analysis of the texture descriptors on a common background.

Local Binary Pattern Histogram Fourier features (LBP-HF), which are a novel rotation invariant image descriptor computed from discrete Fourier transforms of local binary pattern (LBP) histograms, were proposed in collaboration with Prof. Jiri Matas. Unlike most other histogram based invariant texture descriptors which normalize rotation locally, the proposed invariants are computed globally for the whole region. In addition to being rotation invariant, the LBP-HF features retain the highly discriminative nature of LBP histograms. In the experiments, it was shown that these features outperform non-invariant and earlier version of rotation invariant LBP and the MR8 descriptor in texture classification, material categorization and face recognition tests.

We continued research on a novel local descriptor based on Weber’s law (WLD), originally developed by our postdoctoral researcher Dr. Jie Chen and his colleagues at the Institute of Computing Technology of the Chinese Academy of Sciences. A CVPR paper published on this topic in 2008 was extended to a journal article, for example by improving the theoretical part of the paper, developing a multiscale WLD operator, and carrying out new experiments on texture classification.

Since 2005, we have been investigating methods for analyzing dynamic textures (DT), i.e. textures in motion. A spatiotemporal volume LBP operator (VLBP) was developed which combines temporal and spatial information (i.e. motion and appearance). A simplified method based on concatenated LBP histograms computed from three orthogonal planes (LBP-TOP) was proposed.

In the reporting year, we addressed the problem of segmenting DT into disjoint regions in an unsupervised way. Each region is characterized by histograms of local binary patterns and contrast in a spatiotemporal mode, combining the motion and appearance of DT together. Experimental results showed that our method is effective in segmenting regions that differ in their dynamics. Experimental results
and a method with some existing methods showed that our method is effective for DT segmentation, and is also computationally simple compared with methods such as those using mixtures of a dynamic texture model or level sets. Furthermore, our method performs fairly well on sequences with clustered background.

We also launched new research on video texture synthesis using spatiotemporal local binary patterns. Dynamic texture synthesis is to provide a continuous and infinitely varying stream of images by doing operations on dynamic textures. We proposed a novel frame-feature descriptor accompanied by a similarity measure using a spatial-temporal descriptor: LBP-TOP which considers both the spatial and temporal domains of video sequences; moreover, it combines the local and global description on each spatiotemporal plane. The preliminary results are very promising.

We have adopted spatiotemporal LBPs to the problems of facial expression recognition, visual speech recognition, recognition of actions, and gait recognition. These topics will be discussed in the section Human-Centered Vision Systems.

We have also begun new collaboration with Docent (Adjunct Professor) Johan Lundin from the Biomedical Informatics Group at the University of Helsinki. The aim of this research is to apply texture based methods for breast cancer cell identification and grade classification using images obtained with their web based virtual microscopy system.

**Geometric Image and Video Analysis**

Imaging geometry has a central role in computer vision as it provides the mathematical foundation for analyzing both 2D and 3D spatial relations from images and video. Our research on geometric image and video analysis aims for the development of basic methodology for imaging geometry related computer vision problems. The topics that have been investigated during 2008 are quasi-dense matching with applications to object recognition and non-rigid image registration, uncertainty analysis of multilinear geometric entities, and blur invariant pattern recognition and matching.

Finding correspondences is a classical problem in computer vision, where one has to determine which pixels in different images correspond to the same 3D point. This is basically an image registration task, and we have studied a quasi-dense matching technique for solving it. Quasi-dense matching is a novel approach based on a match propagation principle which might be useful for various applications. One such application is the recognition of specific objects from unknown test images. This kind of recognition problem can be seen as an image registration problem where the extent of the matching region has to be determined together with the geometric transformation. In this manner we have been able to improve the recognition result using quasi-dense matching compared to such approaches that use only a sparse set of matching regions for recognition. Our approach additionally gives the segmentation of the object. Moreover, in our recent research we have observed that the quasi-dense approach can be useful in difficult motion segmentation problems where the scene is non-rigid.

By non-rigid quasi-dense matching the objects can be recognized and segmented even if their viewpoint or background changes or they are only partially visible.

In our recent research on statistical methods in geometric computer vision, the focus has been on uncertainty analysis of multilinear geometric entities, which include the most usual projective-geometric relationships in computer vision. Especially the theory of dual distributions and integral geometry has been considered. In the research, it evolved that differential and integral geometry, together with the study of Grassmann algebras and projective subspaces provide a promising system for the uncertainty analysis in geometric computer vision.

In many machine vision applications, captured images contain blur due to motion of the camera or the lens being out of focus. We have continued our research on blur invariant pattern recognition and image registration using invariant features based on the Fourier transform phase. These features are invariant to centrally symmetric blur, including linear motion and out of focus blur. The latest development has been a method for combined blur and affine invariant object recognition. Affine transform is a good approximation of realistic perspective transform of images due to a viewpoint change.

We have also developed a completely new method for local image analysis called the local phase quantization (LPQ) descriptor. It utilizes the quantized phase of the local Fourier transform, and is also quite robust to the above mentioned blurring. The descriptor has been used for statistical analysis of local image characteristics in texture classification. The descriptor was already successfully applied for face recognition in the case of slightly blurry images. The latest step has been the development of rotation invariant version of the descriptor.
In many practical situations, it is often desirable to enrich
the information contained in one image by assigning extra
values which encode the relative distance from the observer
to each considered pixel. Efficient recovery of depth val-
ues requires extra information, which is usually provided
in the form of two or more images representing the same
scene captured from different angles. However, estimating
depth information under non-ideal conditions has not been
widely addressed in the literature. We have developed an
efficient and fast method which estimates a depth-map from
a stereo-pair of images impaired by typical radiometric deg-
gradations, such as out-of-focus blur, motion blur and expo-
sure changes. The algorithm exploits the properties of the
LPQ in order to obtain invariance to the aforementioned
types of degradation, and does not rely on a-priori knowl-
dege of radiometric changes.

Human-Centered Vision Systems

It is widely predicted that computing will be moving to the
background, being omnipresent and invisible, and project-
ing the human user into the foreground. Therefore, future
ubiquitous computing environments should be designed as
human-centered instead of computer-centered. Computer
vision will play a key role in implementing human-centered
systems, for example, for natural human-computer interac-
tion (HCI) or for identifying humans and their behavior in
smart environments.

Face Analysis

In 2004, we proposed a novel facial representation based
on LBP features, obtaining excellent results. A paper on
this topic was published in 2006 in IEEE Transactions on
Pattern Analysis and Machine Intelligence. Our approach
has evolved to be a growing success. It has been adopted
and further developed by many research groups and lead-
ing scientists working in the field.

Differences in illumination conditions cause significant
challenges for any 2D face recognition algorithm. One of
the methods to counter these effects is image preprocessing
before the feature extraction. We proposed a new prepro-
cessing approach that uses custom filters obtained through
an optimization procedure striving for most suitable pre-
processing filters for the selected feature extractor and dis-
tance measure. We experimented with it using Local Binary
Pattern texture features and Chi-square histogram distance
metric. Results obtained with the Face Recognition Grand
Challenge (FRGC) 1.0.4 dataset showed that our method
performs better than perhaps the best existing preprocessing
method recently proposed by Tan and Triggs.

LBP histograms have been successfully used in face detec-
tion, recognition, verification, facial expression recognition
etc. The models for face description have been based on
LBP histograms computed within small image blocks. Re-
cently, we developed a novel, spatially more precise model,
based on kernel density estimation of local LBP distribu-
tions. Our experiments showed that this model produces
significantly better performance in face verification tasks
than the earlier models. Furthermore, we proposed and
evaluated the use of a Support Vector Machine (SVM) in
information fusion from individual pixels for the binary clas-
sification task for identity verification.

Research on facial expression recognition was continued.
Facial expressions can be thought of as specific dynamic
textures where local appearance and motion information
need to be taken into account. We utilize local spatiotem-
poral LBP-TOP operators to describe facial expressions.
All existing facial expression recognition databases are cap-
tured in a visible light spectrum. Visible light usually changes
with locations, and can also vary with time, which can cause
significant variations in image appearance and texture. Near-
infrared (NIR) imaging, on the other hand, provides robust-
ness with respect to illumination changes. We collected a
novel NIR facial expression database to be used in our re-
search. An improved method for facial expression recogni-
tion utilizing multiresolution LBP-TOP features and feature
selection was developed, providing very promising results.
A novel weight based method was also proposed to further
improve the recognition accuracy. A near real-time exper-
imental system was implemented to demonstrate the appli-
cability of our approach.

While much work considers moving faces only as collec-
tions of frames and applies still image based methods, re-
cent developments indicate that outstanding results can be
obtained using texture based spatiotemporal representations
for describing and analyzing faces in videos. Such scenarios
are commonly encountered in many applications such as
human-computer interaction and visual surveillance in
which input data generally consists of video sequences. In-
spired by psychophysical findings which state that facial
movements can provide valuable information for face analy-
sis, and also inspired by our recent success of using LBP
for combining appearance and motion for dynamic texture
analysis, we investigated the combination of appearance
(the shape of the face) and motion (the way a person is
talking and moving his/her facial features) for face analysis
in videos. We proposed and studied an approach for spa-
tiotemporal face and gender recognition from videos using
an extended set of Volume LBP features and a boosting
scheme. We experimented with several publicly available
video face databases, and considered different benchmark
methods for comparison. Our extensive experimental analy-
sis clearly assessed the promising performance of the LBP
based spatiotemporal representations for describing and
analyzing faces in videos.
We also proposed another approach to gender recognition for cases where face sequences are available. Instead of treating each facial image as an isolated pattern, and then combining the results (at feature, decision or score levels), as is generally done in previous work, we exploit the correlation between the face images and look at the problem of gender classification from a manifold learning point of view. Our approach consists of first learning and discovering the hidden low-dimensional structure of male and female manifolds using an extension to the Locally Linear Embedding algorithm. Then, a target face sequence is projected into both manifolds for determining the gender of the person in the sequence. The matching is achieved using a new manifold distance measure. Extensive experiments on a large set of face sequences and different image resolutions showed very promising results, outperforming many traditional approaches.

In 2008, we have also started investigating a new research area called soft biometrics. In contrast to "hard" biometrics, which includes face, fingerprint, retina, iris, voice etc., and are generally unique and permanent personal characteristics, soft biometrics (including age, beard, gender, glasses, ethnicity, eye/hair color, height/weight, skin color etc.) provide some vague physical or behavioral information which is not necessarily permanent or distinctive. This is very useful in many applications such as human-machine interaction, and content based image/video retrieval. Our very preliminary experiments on age and ethnic classification problems showed promising results.

We also continued our research on recognizing isolated phrases using only visual information. We use spatiotemporal LBP-TOP descriptors extracted from mouth regions to represent and recognize spoken phrases. Our approach has outperformed earlier methods, for example in experiments with the well-known AVLetters database. The advantages of our approach include local processing and robustness to monotonic gray-scale changes. Moreover, no error prone segmentation of moving lips is needed. We began research on continuous speech recognition, obtaining promising preliminary results.

In the period 2008-2010, the Machine Vision Group is participating in the Mobile Biometry (MOBIO) project funded by the European Commission. The scientific and technical objectives of the project include robust-to-illumination face authentication, robust-to-noise speaker authentication, joint bi-modal authentication, model adaptation and scalability. The LBP method developed in MVG plays an important role in the project. In the first year of the project, we actively participated in the design and collection of a multibiometric research database that is recorded using mobile phones during the course of the project, and at later stages used for the development and evaluation of mobile biometric systems. Furthermore, MVG delivered two baseline face detector systems to the project and organized a project meeting gathering 11 attendees from the participating universities and research institutes to Oulu.

In collaboration with Prof. Stan Z. Li and his students from the Chinese Academy of Sciences, we investigated a new problem in face recognition research in which the face samples for enrollment and query are captured under different lighting conditions. In our case, the enrollment samples are visual light (VIS) images, whereas the query samples are taken in near infrared (NIR) conditions. It is very difficult to directly match the face samples captured in these two lighting conditions due to the different visual appearances of VIS and NIR images. We proposed a novel method for synthesizing VIS face images from NIR images, based on learning the mappings between images of different spectra (i.e., NIR and VIS) images. In our approach, we reduce the inter-spectral differences significantly, thus allowing effective matching between faces taken in different imaging conditions. Face recognition experiments on a data set of 250 subjects clearly showed the efficacy of the proposed approach.
Human-Computer Interaction

We have continued our research on vision based human-computer interaction. Such technologies are likely to be the building blocks for the cognitive systems embedded in homes, offices, vehicles, and the equipment we use for everyday tasks.

Navigating large information spaces can be disorienting, even on a large screen. In mobile devices with a small screen, the user often encounters situations where the content that is needed for display exceeds what can be shown on the screen. For example, large digital images are becoming commonplace due to the increasing availability of high resolution imaging and map navigation applications. A viable alternative for improving interaction capabilities is spatially aware displays. The solution is to provide a window on a larger virtual workspace where the user can access more information by moving the device around.

For this purpose, we have developed a method for 3D face tracking that can be used to control spatially aware user interfaces of mobile devices. Unlike many other methods proposed in the literature, the low computational cost of our method makes it practical for mobile platforms where high computational resources are not available. We also wish to emphasize the point here that our application differs from the usual case since the device is moved with respect to the face. The proposed system consists of two stages. In the initialization stage, the user’s face and eyes are detected automatically. During tracking, an extended Kalman filter estimates the camera pose utilizing a novel combination of motion features and eye positions detected from the face region. Experiments show that the method enables position based control of mobile user interfaces.

Analysis of Motion and Human Actions

Our work on the recognition of human body parts from silhouette images based on statistical models has progressed toward application feasibility. Unlike many other methods proposed in the literature, the low computational cost of our method makes it practical for mobile platforms where high computational resources are not available. We also wish to emphasize the point here that our application differs from the usual case since the device is moved with respect to the face. The proposed system consists of two stages. In the initialization stage, the user’s face and eyes are detected automatically. During tracking, an extended Kalman filter estimates the camera pose utilizing a novel combination of motion features and eye positions detected from the face region. Experiments show that the method enables position based control of mobile user interfaces.

We also developed an approach for human gait recognition that inherently combines appearance and motion. The LBP-TOP descriptors are used to describe human gait in a spatiotemporal way. We proposed new coding of multiresolution uniform local binary patterns, and used it in the construction of spatiotemporal LBP histograms. We showed the suitability of this representation for gait recognition and tested our method on the popular CMU MoBo dataset, obtaining excellent results in comparison to the state of the art methods.

Moving object detection and tracking can be done using different features like statistical color and texture descriptors. They can also be used for tracking objects from one camera view to another to build up a large scale picture of object’s motion and behavior. In collaboration with the Chinese Academy of Sciences (Prof. Stan Z. Li, Institute of Automation), we have continued our work on moving object detection and tracking, and have been investigating different feature boosting and classification techniques for feature selection in view-to-view multi-camera tracking. It should be possible to use these techniques together with the open sensor network framework which was also developed earlier in 2008.
Vision Systems Engineering

To enable useful real-world systems, our vision system engineering research provides guidelines for methodological research, helping to identify attractive approaches, architectures, and algorithms, as general purpose computing is seldom a realistic option. In practice, solutions from low-level image processing to even equipment installation and operating procedures need to be considered simultaneously. The roots of our expertise lie in our industrial visual inspection research in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. Recently, we have applied our expertise to applications intended for smart environments, mobile platforms, and our collaborative vision computing architecture research is a recent spin-off.

We have introduced an open and extendable framework for the development of distributed sensor networks with an emphasis on peer-to-peer networking. The user is provided with easy access to sensors and communication channels between distributed nodes, allowing the effort to be focused on the development of machine vision algorithms and their use in distributed environments.

A demonstration system was implemented to test the suitability of the framework for a distributed multi-camera application. The system is composed of a set of processing nodes running on PC workstations, and Axis 210A/213 IP-cameras acting as sensors. The processing nodes receive image data from the cameras through the sensor interface and detect humans in the current view. Every time a node detects an old or new object, it informs the other nodes through the network interface. Object data and features from the detector are sent to other nodes using XML with an encoded thumbnail image of the detected object. The system also includes a specialized UI node that collects and visualizes all the messages sent by the processing nodes.

We have also started to develop a vision based system for detecting humans from moving work machines. The aim is to combine several detection techniques in order to achieve reliable results in various imaging conditions. The system constructed has both narrow-field and fisheye stereo cameras for capturing the image data, which makes it possible to use the same system also for image based 3D modeling of the work machine environment. Several human detection methods using 2D pattern recognition, motion analysis, and 3D reconstruction have already been implemented and experimented with. The results indicate that all methods have their weaknesses in practical situations, and therefore, a fusion of several methods should be used.

Detecting humans from moving work machines in various imaging conditions is challenging.

An example of a novel application intended for mobile platforms is the panorama imager that “glues” together frames selected from video sequences. The selection process analyzes the displacements between the video frames, measures the blur due to motion and focusing, selecting the suitable frames for mosaicing, and detects moving targets and human faces ensuring that they are not mutilated in the process. In other words, the apparently simple panorama capturing process contains lots of image analysis functionality to achieve good image quality. Some of the solutions developed for mobile platforms are already being re-used in industrial applications. For instance, the frame selection techniques of panorama capture are employed in developing a matrix camera based quality monitoring system for a printing machine that can cope with flutter and frequent environmental disturbances. Additional uses have been found in applications that have previously employed line-scan cameras.

A distributed camera network is set up at the Information Processing Laboratory.
Machine vision applications are characterized by both high data input rates and high computational costs. For instance, a typical raw digital video rate is around 10 Mpixels/s, and its processing demands are at least a few hundred operations per pixel, requiring multiple GOPs of computing power. Often this needs to be done in a small package, such as a mobile communications device that may allocate at most 500 mW for application processing like video coding or person identification. This prevents implementations based on conventional processors even in the future, but hardware acceleration is mandatory. Currently, only monolithic long latency accelerators improve energy efficiency, but they are rigid and costly to design and verify, and difficult to justify for purposes which are considered marginal and computationally expensive.

Together with the Tampere University of Technology (Prof. Jarmo Takala), the Helsinki University of Technology (Prof. Petri Vuorimaa), and the Åbo Akademi University (Prof. Johan Lilium) we are concentrating on improving the energy efficiency of embedded high performance computing. We have demonstrated that fine grained, silicon area efficient adaptable hardware accelerators can be employed, at very low software interface overheads through deterministic multithreaded schedules. This has been shown to be much more efficient than the conventional interrupt, semaphore, and the event handler mechanisms advocated by the textbooks. In essence, we are targeting a new paradigm for embedded computing and expect significant impacts in the field.

Our current demonstrations include the simultaneous decoding of multiple MPEG-4 streams on shared accelerators and MIMO reception. This work has been carried out in cooperation with Centre for Wireless Communications (CWC) and, so far, within application specific processors. More recently, the enabling potential of this technology has been understood in the context of very low cost zero-power devices that have in part been manufactured with printing technology.

Exploitation of Results

Our approach of combining world-class basic research with more applied research on vision systems and systems engineering is quite unique, giving rise to our research having a great practical impact. We conceive of machine vision research as a remarkable field of science that contributes to the competitiveness of Finnish enterprises by developing new technologies. Typical inspection targets include part assemblies in the electronics and car industry, continuous webs such as paper, steel and fabrics, and natural materials such as wooden boards and coffee beans. Many of these targets are textured and colored, such as wood, and the inspection problem is solved best with the respective texture and color based methods. Several inspection systems based on our results are being marketed by our industrial partners for applications ranging from coffee bean sorting to paper formation measurement.

Currently, we are investigating methods and means for building visual inspection systems for exceptionally demanding applications such as non-destructive, non-contact dynamic strength grading of wooden boards. The underlying observation is that the strength is different to the axial, tangential and radial directions of the wood grain, while the behavior of the grain is affected differently by sound and dry knots. After color and texture based image analysis has provided the grain and knot information, a 3D Finite Element Model (FEM) is built and analyzed for strength.

Our work is already internationally noted and exploited. Mr. Jani Boutellier from our group visited the Processor Architecture Laboratory of the École Polytechnique Fédérale de Lausanne (EPFL) and participated in the development of a scheduler for the emerging ISO standard of Reconfigurable Video Coding, as well as in the development of a scheduler hardware circuit. The methodology used is essentially the same as that used in our research, developed in cooperation with Prof. Shuvra Bhattacharyya at the University of Maryland, where Mr. Boutellier was a visitor during the fall of 2006. These contributions are expected to be of significant practical importance in industry when improved design tools supporting the new approach are available.

In 2009, a three year Academy of Finland funded project PARadigm Shift for Embedded Computing (PARSEC) will begin, with a view to automating the tools for designing ultra-energy-efficient systems for computationally demanding purposes, such as video coding and image analysis. The research is being carried out in cooperation with Tampere University of Technology (Prof. Jarmo Takala) and Åbo Academi University (Prof. Johan Lilius).

Visual inspection is economically still the most important application area of machine vision. The inspection systems can be relatively expensive as long as they provide high added value, and are therefore attractive testing grounds for new technologies. Typical inspection targets include part assemblies in the electronics and car industry, continuous webs such as paper, steel and fabrics, and natural materials such as wooden boards and coffee beans. Many of these targets are textured and colored, such as wood, and the inspection problem is solved best with the respective texture and color based methods. Several inspection systems based on our results are being marketed by our industrial partners for applications ranging from coffee bean sorting to paper formation measurement.
problems. The solutions synthesize infrared imaging, human action recognition, and visual learning technologies that make the systems installable in almost any home. The first uses of the technology are expected to be in retirement and nursing homes.

Another example of the impact of our work is that in 2005 Intopii Ltd., a spin-out company from our texture research, entered into a cooperative agreement with the Cognex Corporation, the world’s leading supplier of machine vision systems. In late 2006, another spin-out, Visidon Ltd., was launched. The company provides intelligent computer vision solutions for mobile devices, as well as special system, algorithm, and software design and training services for a variety of demanding industrial and consumer applications.

Future Goals

The recent years have been very successful for the Machine Vision Group. Working as a single well-focused research group, in which different teams and researchers work closely together, has made it possible for them to benefit from each other’s work and cumulative past experiences in an efficient way. Our global network has been strengthening all the time and the collaboration with selected international partners is rather intense. We expect to have new offers for collaboration in the future as well.

During the past two years the group has been able to receive funding for several projects from the Academy of Finland, which will guarantee that the share of our long-term basic research will remain significant.

In January 2009, a joint research effort between the Machine Vision Group and the Intelligent Systems Group will be launched. The Academy of Finland will fund this new four-year research project on affective human-robot interaction via its “Ubiquitous computing and diversity of communication” research program. The project combines the firm expertise of the both groups. The Machine Vision Group has reached very promising results in automatic recognition of facial and body expressions, and even speech from videos. The Intelligent Systems Group has wide expertise in designing mobile robot platforms and software for their control and behavior.

Development of affective human-computer interfaces (HCI) is of great interest in building future ubiquitous computing (Ubicom) systems. Within the next 15 years, domestic servicing robots can replace the human in many routine tasks in our everyday life. Human-robot interaction will take place locally in a “face to face” manner, as well as remotely using a mobile device and wireless communication.

The project aims to produce leading-edge approaches for affective human-robot interaction in smart Ubicom environments. An intelligent robot will detect and identify the user, and personalize and customize its services according to this information. It will recognize the emotions of the user. The communication between the human and the robot will be natural since the robot can understand commands given by the human through speech or gestures.

Personnel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors &amp; doctors</td>
<td>10</td>
</tr>
<tr>
<td>graduate students</td>
<td>16</td>
</tr>
<tr>
<td>others</td>
<td>15</td>
</tr>
<tr>
<td>total</td>
<td>43</td>
</tr>
<tr>
<td>person years</td>
<td>27</td>
</tr>
</tbody>
</table>

External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>385 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>201 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>274 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>184 000</td>
</tr>
<tr>
<td>international</td>
<td>116 000</td>
</tr>
<tr>
<td>total</td>
<td>1 160 000</td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


Background and Mission

The Machine Vision Group (MVG) is renowned world-wide for its expertise in computer vision. The group has a solid record, which now spans 28 years, of scientific merits on both basic and applied research on computer vision. The mission of the group is to always pursue actual research challenges and improve the state-of-the-art methods.

MVG works as a single well-focused research group, and intensively collaborates with other groups of similar status in Europe, the USA, China and Japan. The group has been invited to take part in several European project proposals, and during the reporting period had a joint research project on face analysis and visual surveillance together with Prof. Stan Z. Li from the Institute of Automation at the Chinese Academy of Sciences. These are clear signals of our attractiveness as a distinguished partner in the global research community. Within the Seventh Framework Programme FP7, the group currently participates in a project consortium of Mobile Biometry (MOBIO), which is coordinated by the IDIAP Research Institute, Switzerland. A new project called Tabula Rasa will continue the biometry theme and the MVG has a significant role in this as well.

The research areas of the group range from generic computer vision methodologies to machine vision applications and vision systems engineering. The results of our research have been widely exploited in industry, and contract research forms a part of our activities. The current areas of research interest include texture analysis, geometric computer vision, computational imaging, face analysis, analysis of motion and human actions, vision-based human-computer interaction and energy-efficient architectures for vision computing.

Highlights and Events in 2009

During 2009, the group produced a large number of significant scientific contributions and published its key results through major conferences and journals. The group has been recently very successful in winning highly competitive funding for its long term research from the Academy of Finland. Altogether, as many as eight Academy-funded research or postdoctoral projects were running in 2009. As a highlight, the Affective Human-Robot Interaction project aims at creating leading edge solutions for natural human-robot interaction in smart environments.

The research group has established its position among the most significant research teams in the field of machine vision. One of the indicators of this is, that the Local Binary Pattern (LBP) papers of the group have received over 3 300 citations in Google Scholar by the first quarter of 2010. An invited article on the LBP methodology has appeared on Scholarpedia: www.scholarpedia.org/article/Local_Binary_Pattern.

In addition, visibility outside academic forums has been achieved. The group has a tradition of presenting regularly its research projects and activities in different media. Last year, this included, for example, coverage in the leading Finnish newspaper Helsingin Sanomat, a national professional magazine Prosessori, and the local newspaper Kaleva.

The group hosts regularly visits by respected and renowned scientists from abroad. In 2009, the group had the pleasure of hosting Professor Hanan Samet from the University of Maryland, USA and Prof. Georgy Gimel’farb from the University of Auckland, New Zealand. In addition, several domestic visitors from both public and private sectors were briefed about our research activities.

The group has established active collaboration with some of the world’s leading institutions and top scientists. It has had collaboration with the University of Maryland (USA) from the early 1980’s. More recent partners include the Chinese Academy of Sciences, the Academy of Sciences of the Czech Republic, Czech Technical University in Prague, the École Polytechnique Fédérale de Lausanne (EPFL), and the consortium of the European Mobile Biometry (MOBIO) project. In 2009, MVG established contacts with the University of Tokyo, Japan and the University of Oxford, UK with postdoctoral researcher visits.

In 2009 MVG hosted visitors from the Chinese Academy of Sciences.

The group fosters international mobility to and from our unit. Four of our researchers made research visits to foreign institutions during the reporting period. The group has at-
Texture is a fundamental property of surfaces. We have de-
tmented visiting postdoctoral researchers and postgraduate
students from abroad, who are affiliated with us for periods
of a couple of weeks up to several years. Three such short
visits were made to MVG in 2009 and one long visit began.
In addition, MVG recruited as many as three new post doc-
toral scientists from abroad.

The group and its members are active in the scientific com-

The group has a long and highly successful research tradi-
tion in two important generic areas of computer vision: tex-
ture analysis, and geometric computer vision. The
MVG presented tutorials at two major conferences of the field: the IEEE International
Conference on Computer Vision in Kyoto, Japan (ICCV 2009)
and the Asian Conference on Computer Vision in Xi’an,
China (ACCV 2009). Prof. Matti Pietikäinen and Dr. Guoying
Zhao hosted a tutorial on “Local Texture Descriptors in
Computer Vision” at ICCV and Dr. Abdenour Hadid a tutor-
ial on “Image and Video Analysis using Local Binary Pat-
terns (LBP)” at ACCV.

The professors of the group were committee members of
several major conferences and many researchers of the
group served as reviewers for various journal and confer-
ence articles. Dr. Guoying Zhao served as a co-editor for a
book entitled “Machine learning for human motion analy-
sis: theory and practice”, IGI Global, 2009, together with Dr.
Liang Wang and Dr. Li Cheng.

Dr. Guoying Zhao has also been serving as a guest editor of the
special issue New Advances in Video-based Gait Analy-

Scientific Progress

The current main areas of the research are: 1) Computer
vision methods, 2) Human-centered vision systems, and 3) Vision
systems engineering.

Computer Vision Methods

The group has a long and highly successful research tradi-
tion in two important generic areas of computer vision: tex-
ture analysis, and geometric computer vision. The
fundamental research in these areas has created a basis for
many novel contributions in our research on human-cen-
tered vision systems and vision systems engineering. In
2009, we also started a new research area, computational
imaging, which has recently gained much attention within
the computer vision and computer graphics communities.

Texture Analysis

Texture is a fundamental property of surfaces. We have de-
veloped a novel methodology based on Local Binary Pat-
terns (LBP), which has evolved to represent a breakthrough
in texture analysis. Today, LBP is probably the most widely
used texture operator. It is frequently cited and used all over
the world. With over 430 citations in early 2010, our paper
on local binary patterns published in IEEE Transactions on
Pattern Analysis and Machine Intelligence (PAMI 2002) is,
according to the ISI Web of Knowledge, the most cited
Finnish paper published since 2002 e.g. in electrical & elec-
tronic engineering, computer science and telecommunications
categories, and the 8th most cited paper of the
prestigious PAMI journal. The group’s more recent LBP-
based publications on face recognition (PAMI 2006) and
dynamic texture and facial expression recognition (PAMI
2007) were the 3rd and 4th most cited papers of PAMI
published since 2006 and 2007, respectively.

In 2009, our research focused on robust LBP based descrip-
tors for static and dynamic textures, and on texture based
methods for different tasks in face and activity analysis.

The purpose of Timo Ahonen’s doctoral thesis “Face and
texture image analysis with quantized filter response statistics”, published in August 2009, was the development and
analysis of photometric descriptors for the appearance of
real life images. The two application areas included in the
thesis were face recognition and texture classification. The
contributions of the thesis include a face representation
method using distributions of local binary patterns com-
puted in local rectangular regions, a face recognition method
tolerant to image blur using local phase quantization, and
three new approaches and extensions to texture analysis
based on a general framework for image description using
statistics of quantized filter bank responses. The first two
aim at increasing the robustness of the quantization pro-
cess. The soft local binary pattern operator accomplishes
this by making a soft quantization to several labels, whereas
Bayesian local binary patterns make use of a prior distribu-
tion of labeling, and aim for the one maximizing the a poste-
riori probability. Third, a novel method for computing rotation
invariant statistics from histograms of local binary pattern
labels using the discrete Fourier transform was introduced.

In April 2010, Ahonen’s thesis was ranked by The Pattern
Recognition Society of Finland as the best Finnish PhD

In breast cancer classification, the dominant factor is the
histological appearance of the tissue. In collaboration with
Docent Johan Lundin from the Biomedical Informatics Group
at the University of Helsinki, we have been developing an
automatic stroma-parenchyma tissue classification system.

Automatic tissue classification could speed up, focus and
add reliability to the visual inspection of histopathological
images. Our experimental results show that joint distribu-
tion of Local Binary Pattern codes and local contrast (LBP/
C) is a strong texture descriptor for histopathological breast
images. Very promising results were achieved using the
AdaBoost algorithm and Support Vector Machine (SVM)
classifier with LBP/C features. Related to this topic, we are
also collaborating in a mobile microscopy project, which
aims to develop a device for medical diagnostics particu-
larly used in developing countries.
In collaboration with the Institute of Computing Technology of the Chinese Academy of Sciences, we have proposed a simple, yet very powerful and robust local descriptor called the Weber Law Descriptor (WLD). It is based on the fact that human perception of a pattern depends not only on the change of a stimulus (such as sound, lighting) but also on the original intensity of the stimulus. Specifically, WLD consists of two components: differential excitation and orientation. The differential excitation component is a function of the ratio between two terms: one is the relative intensity differences of a current pixel against its neighbors; the other is the intensity of the current pixel. The orientation component is the gradient orientation of the current pixel. For a given image, we use the two components to construct a concatenated WLD histogram. Experimental results on texture analysis and face detection problems provided excellent performance. A paper on this topic was accepted by the top-tier IEEE Transactions on Pattern Analysis and Machine Intelligence journal.

Dynamic textures (DT) are image sequences with visual pattern repetition in time and space, such as smoke, flames, moving objects and so on. The Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) method which we recently proposed is an effective descriptor for dynamic texture recognition, as shown by its success in different application problems. A problem with the LBP-TOP descriptor is that it is not rotation invariant, which makes it difficult to deal with rotation variations common in video sequences. Inspired by Local Binary Pattern Histogram Fourier features (LBP-HF), which are a novel rotation invariant image descriptor computed from discrete Fourier transforms of local binary pattern (LBP) histograms, research on extending it to acquire rotation invariant LBP-TOP descriptors was carried out with very good preliminary results.

How to segment dynamic textures is a challenging problem. In 2009, we developed significant improvements to our DT segmentation method published at ICPR 2008. We employ a new spatiotemporal local texture descriptor which combines local binary patterns with a differential excitation measure of the WLD descriptor. We also addressed the important problem of threshold selection by proposing a method for determining thresholds for the segmentation method by statistical learning. An improved criterion for merging adjacent regions was also introduced. Experimental results show that our approach provides very good segmentation results compared to the state-of-the-art.

Dynamic texture synthesis is to provide a continuous and infinitely varying stream of images by performing operations on dynamic textures. The starting point for our research was that the previous video texture method proposed by Schödl et al. (SIGGRAPH 2000) provides good quality visual results, but does not well explore the temporal correlation among frames. As an improvement, we developed a novel spatial temporal descriptor for frame description accompanied with a similarity measure. Our method considers both the spatial and temporal domains of video sequences in representation; moreover, it combines the local and global description on each spatial-temporal plane. To the best of our knowledge, this is the first utilization of texture properties in both the spatial and temporal domains in frame representation for dynamic texture synthesis. A paper on this topic was published in the International Conference on Image Processing (ICIP).

We have continued our work on the novel Local Phase Quantization (LPQ) descriptor which is based on a locally...
computed Fourier transform phase. LPQ is insensitive to image blurring, and it has proven to be a very efficient descriptor in face recognition from blurred as well as sharp images. In our latest research, we have compared different approaches for computing the local phase information. According to our experiments, we have found that ordinary short-term Fourier transform can provide the best results in the case of blurred images. We have also improved the blur robustness of the descriptor by introducing an optimization framework that can adjust the descriptor to better tolerate different types of blur, including motion, defocus and Gaussian blur.

**Geometric Computer Vision**

Geometric camera calibration has traditionally been one of our areas of expertise. Recently, our contributions in this research area were summarized in a comprehensive review article, which was published in the Wiley Encyclopedia of Computer Science and Engineering in 2009. The review article gives an overview of camera models and calibration methods used in the field. The emphasis is on conventional calibration methods in which the parameters of the camera model are determined by using images of a calibration object whose geometric properties are known. In addition, the presented techniques are illustrated with real calibration examples in which various kinds of cameras, both omnidirectional and conventional ones, are calibrated using open source software tools developed in our group.

One recent trend in many vision problems which involve image matching, is the success of correspondence growing methods. For example, in object recognition and image retrieval, the idea of correspondence growing has been used to improve the discrimination between correct and incorrect region correspondences. On the other hand, in multi-view reconstruction, patch-based approaches, which use local match expansion, have produced good results. Our previously proposed quasi-dense wide baseline matching method is one particular growing technique which can be used as a basic tool in different recognition and reconstruction tasks. During the year 2009, we have continued our studies with quasi-dense matching and its applications. One of the recent developments is a dense and deformable motion segmentation method for wide baseline image pairs. In this work, we proposed a bottom-up motion segmentation approach which gradually expands and merges the initial matching regions into smooth motion layers and finally provides a dense assignment of pixels into these layers. Besides segmentation, the proposed method provides the geometric and photometric transformations for each layer. Hence, due to its ability to deal with deforming motions and large illumination changes, our approach has a wider range of applicability than the previous methods.

In many computer vision applications, images are degraded by motion or defocus blur, which may cause significant problems to image analysis. To overcome these problems, one can use deblurring techniques to eliminate the blur, but this is likely to result in disturbing artefacts. Another approach is to employ such features in image analysis that are insensitive to blur. In 2009, we have continued our research on blur invariant pattern recognition and registration. We have developed global image features that are based on the Fourier transform phase. In contrast to the LPQ, these features are global descriptors that can be used to geometrically align images and recognize objects that are geometrically transformed. Their main advantage is that they are invariant to centrally symmetric blur kernels including linear motion and out of focus blur. In our latest work, the results obtained with these invariants were greatly improved by using a statistical method to weight the features according to their estimated signal-to-noise ratio. When this method was applied to recognition of blurred and noisy objects, the results were improved by up to 20%.

**Computational Imaging**

Computational imaging is a relatively new field of research, and it refers to the convergence of the camera and the computer. Computational cameras are able to enhance and extend the capabilities of ordinary photographic cameras by using computational techniques to process and reconstruct the image data provided by the camera sensor. In our research, we have developed methods to process the raw sensor data in order to improve the subjective quality of the resulting images by eliminating or correcting the effects of certain error sources.

Color constancy or white balance adjustment is a fundamental element in virtually all modern digital cameras. Accurate color constancy is crucial, since even small errors may cause severe deterioration in the empirical quality of the output image. Current methods for color constancy rely on general assumptions of image colors statistics (e.g. the mean color of the image is always gray), which further require information about the sensor characteristics. Although such approaches work adequately well with an average photograph, they easily make coarse errors when the preselected assumptions are not fulfilled (e.g. the mean color in an image of football field is far from being gray). The situation is even worse if accurate sensor characteristics are not available, which is commonly the case with low cost sensors.

In our research, we have applied higher level content understanding to color constancy estimation. The developed algorithm finds semantically meaningful objects from the image, whose color is known in advance. If such objects, e.g. grass or snow, are found, we can use this knowledge to make more realistic assumptions on the color statistics and sensor characteristics to be applied in color adjustments.

During the acquisition of a digital image, the camera sensor array records the spectral irradiances values for each pixel location. Raw data of this kind do not usually give a representation of the scene as a human would perceive it. They
must be processed and refined in several stages before a useful image is output to the user. However, the operations performed at each stage also contribute to altering the original properties of the sensor noise.

Since the properties of noise for a specific sensor are usually known, one can obtain a fairly accurate mathematical model which helps to improve noise reduction. This is the main advantage of de-noising raw data. There exist many algorithms for raw image de-noising, but it is not trivial to measure their objective and subjective performance, and the literature lacks exhaustive comparisons between different methods. We proposed a novel quality measure for color images which is based on known mechanism of human visual perception. We then executed an extensive session of experiments with real and artificially degraded noisy images using several state-of-the-art de-noising algorithms, and evaluated their performance with ours and other quality measures proposed in the literature.

**Human-Centered Vision Systems**

It is widely predicted that computing will be moving into the background, being omnipresent and invisible, and projecting the human user into the foreground. Therefore future ubiquitous computing environments should be designed to be human-centered instead of computer-centered. Computer vision will play a key role in implementing human-centered systems, for example, for natural human-computer interaction (HCI) or for identifying humans and their behavior in smart environments.

**Face Analysis**

In 2009, we continued the research on soft biometrics from videos, and introduced new methods based on manifold learning. Soft biometrics, including age, beard, gender, glasses, ethnicity, eye/hair color, height/weight, skin color etc., provide some vague physical or behavioral information which is not necessarily permanent or distinctive. This is very useful in many applications such as human-machine interaction, and content based image/video retrieval. Our extensive experiments on the gender and age classification problems showed that the developed manifold learning based approach yields excellent results, outperforming those of traditional methods. The idea behind the proposed approach consists of learning the intrinsic characteristics of the subjects in each class (e.g., male, female, young etc.) from the training video sequences by discovering the hidden low-dimensional nonlinear manifold of these subjects.

In cooperation with Prof. Yoichi Sato and his team from the University of Tokyo (Japan), we have also started investigating the recognition of blurred faces combining facial deblurring and LPQ, gaze estimation using appearance based methods, and audio-video correlation analysis for spoofing detection in biometric systems.

In the period 2008-2010, the Machine Vision Group is participating in the Mobile Biometry (MOBIO) project funded by the European Commission. The scientific and technical objectives of the project include robust-to-illumination face authentication, robust-to-noise speaker authentication, joint bi-modal authentication, model adaptation and scalability. The LBP method developed in MVG plays an important role in the project. In the second year of the project, we actively participated in the design and collection of a multibiometric research database that is recorded using mobile phones during the course of the project, and at later stages used for the development and evaluation of mobile biometric systems. Furthermore, MVG carried out research on face recognition methods based on local binary patterns and local phase quantization. In this research, we had collaboration with Prof. Josef Kittler’s group from the University of Surrey, UK. We have also been developing a scalable face detector, which will be considered for the final demonstration system using the Nokia N900 mobile phone.

In collaboration with Prof. Stan Li’s group, we have collected a novel Oulu-CASIA NIR-VIS facial expression database, which includes the six facial expressions from both near infrared (NIR) and visible light (VIS) imaging systems with 80 people (50 from Oulu and 30 from CASIA). We have carried out experiments on this database using LBP-TOP for illumination invariant facial expression recognition. We plan to release this database and provide the baseline results.

We have made an extensive investigation on visual speech recognition. Visual speech information plays an important role in lip reading under noisy conditions, or for listeners with a hearing impairment. We utilized local spatiotemporal descriptors to represent and recognize spoken isolated phrases based solely on visual input. Spatiotemporal local binary patterns extracted from mouth regions are used for describing isolated phrase sequences. The advantages of our approach include local processing and robustness to monotonic gray-scale changes. Moreover, no error prone segmentation of moving lips is needed. The results were published in IEEE Transactions on Multimedia.

We also developed a face tracker, which creates an adaptive color histogram model for tracked objects and applies motion estimation, model based tracking, and generic face detection for keeping track of the faces. Multiple faces can be tracked at the same time using a single camera. Occlusion
Human-Computer Interaction

Methods for organizing images automatically based on their content are becoming more and more attractive as the number of photos in personal collections is rapidly increasing, and finding particular images manually from these collections is becoming too laborious. In our research, we have developed a new generic method for organizing images and image categories. While the traditional approach for categorization is to divide the data into disjoint classes or clusters, the basic idea in our approach is to find the intrinsic structure of the image collections, and to utilize this structure to organize images. Our generic structure for describing the connections between categories creates a more natural and versatile way of organizing the images than the commonly used tree-shaped structures.

Human-Robot Interaction

The objective of the ongoing Academy-funded project “Affective Human-Robot Interaction”, carried out jointly with Prof. Juha Röning’s Robotics Group, is to develop leading edge solutions for affective human-robot interaction (HRI) in smart environments. An intelligent robot should be able to detect and identify the user in order to personalize its services and guarantee security. Furthermore, it should recognize the user’s emotions to allow affective interaction, and be able to communicate easily with the user in order to understand given commands by recognizing speech and gestures. The robot should also learn to change its behavior according to the user and his/her emotional state.

The focus in the research during the first year of the project was in developing new methods and tools for computer vision and robot navigation to be used in integrated HRI systems. The design of an experimental HRI system was also begun.

As a part of this project, research aiming at video-realistic speech animation was also begun. An animated talking face (avatar) with realistic mouth movements and facial expressions can provide a natural way for interaction between humans and robots. In the first phase, problems of visual speech analysis were investigated and a new method based on manifold learning was developed for lip reading. Very promising results were obtained.

Analysis of Motion and Human Actions

The research efforts on detection and tracking moving objects in camera networks resulted in the development of a new algorithm able to re-recognize objects across camera views. The algorithm uses adaptive boosting and a multitude of image properties, like color, texture and shape, to learn object templates, which can then be used to track objects from one view to another. Experiments on real-world datasets suggest that tracking objects in a real-time camera network should be feasible.

In crowd scenarios, people often walk in groups. We developed a new solution to the problem of matching groups of people across multiple non-overlapping cameras. Here, we use “group” to refer to a small number of people walking in proximity. Similar to the problem of matching individuals across cameras, matching groups of people also encounters challenges such as variations of illumination conditions, poses and camera parameters. Moreover, people often swap their positions while walking in a group. We proposed using a covariance descriptor in appearance matching of group images. Experimental results demonstrate the effectiveness of the proposed method.
Human motion can be seen as a type of texture pattern. We adopted the ideas of spatiotemporal analysis and the use of local features for motion description by describing human movements with LBP-TOP dynamic texture features. The proposed description is computationally simple and suitable for various applications such as action and gait recognition. We tested our method with popular Weizmann and KTH datasets for action recognition, and an article describing the results has been accepted by the Machine Vision and Applications journal. Gait recognition results on the CMU MoBo database were published in the International Conference on Biometrics (ICB).

The ability to separate informative regions from background clutter is an essential requirement for a practical vision algorithm. Biological systems have developed to be remarkably effective in focusing their visual attention to relevant targets, unlike computer vision where background subtraction is still an unsolved problem.

Background subtraction is commonly approached by detecting moving objects against a nearly static background. Joint research on this kind of approach was carried out with Prof. Stan Z. Li’s group. The LBP operator was extended to a scale invariant local ternary pattern (SILTP) operator, which is effective for handling illumination variations, especially soft cast shadows. Excellent results were reported for different types of video sequences. A paper with the title “Modelling pixel process with scale invariant local patterns for background subtraction in complex scenes” was accepted for the top-tier IEEE Conference on Computer Vision and Pattern Recognition.

While effective in certain scenes, this kind of approach has severe problems when the scenes are dynamic. A recent approach to background subtraction in dynamic scenes is provided by saliency detectors. The motivation for this approach originates from the attention mechanism of the human visual system, which enables us to quickly focus on general salient objects without preceding training. The most recent saliency detection methods are based on phase response of the estimated Fourier transform. Although this approach is easy to compute, it provides rather poor accuracy and few possibilities to adjust the method for the current application.

In our group, we have developed a new saliency detection method which is based on semi-local feature contrast of the image. Without trading off much of the computational simplicity, we achieve far better accuracy than the existing methods, and also provide several possibilities to tune the detector according to the requirements of the application. The developed algorithm is also applicable to video sequences.

We have also developed a new method for tracking multiple objects from image sequences using detector responses as measurements. The method utilizes soft assignment to associate the detection responses to the objects tracked. Due to soft assignment, we are able to cope with inaccurate responses and inter-object occlusions. The method includes a component which combines Kalman filtering and expectation maximization (EM) algorithms to estimate the parameters of the objects tracked and to assign the measurements softly. One of the further benefits of this approach is that neither iterations nor long measurement history are needed.

Vision Systems Engineering

Vision systems engineering provides guidelines for methodological research, helping to identify attractive computing approaches, architectures, and algorithms for useful commercial systems. In practice, solutions from low-level image processing to even equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our industrial visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. We have also applied our expertise to applications intended for smart environments and embedded platforms such as mobile phones.

On mobile platforms, we have developed a document mosaic constructor that stitches together frames selected from video sequences. The user interactive image capturing process analyzes the displacements between video frames, measures the quality of the frames based on the motion blur and focus, and selects the suitable frames for mosaicing. The implementation includes a parallel pipeline that uses the mobile Graphical Processing Unit (GPU) integrated on the platforms, reducing the computation and power needs of the application.

Application flow of the mobile document mosaic constructor.
Another example of an application that takes advantage of GPU processing is the 3-D image reconstruction software developed in collaboration with VTT. The software constructs a surface map from several images of a sinusoidal pattern projected onto a wooden or metallic sample. The method analyzes the displacements between the images, and computes the phase image using an iterative technique. All the required operations of the algorithm are computed in parallel using general purpose computations on GPUs.

We have also developed a vision based system for distinguishing humans from moving work machines. The system constructed had both narrow-field and fisheye stereo cameras for capturing the image data, which made it possible to use the same system also for image-based 3-D modeling of the work machine environment. Several human detection methods using 2-D pattern recognition, motion analysis, and 3-D reconstruction were implemented and experimented with. The system was evaluated with extensive image sequences of authentic work machine environments, and it proved to be feasible.

The results of our recent body part segmentation work have been applied to unusual activity recognition in home environments. The most important event we seek to detect is falling over. The system developed works robustly when the person monitored is partially occluded or there is noise in the video data. We are developing low cost systems that are easy to install. During the year 2010 we shall begin using cameras with very thin optics. The thin optics lower the overall cost of the systems, but also introduce new challenges for the detection. The resolution of such a camera is very low. The work in the future will focus on adopting the new cameras to the system and testing the performance in real environments.

One of the promising uses for visual inspection is automated strength grading for sawn timber. The solution under development employs real-time feature extraction, classification, and the Finite Element Method (FEM) combined into an adaptive learning scheme. The idea is to launch a FEM based analysis whenever the region under inspection deviates from those modeled earlier. The 3-D model and element mesh needed in the FEM analysis are built from the data.
acquired during feature extraction, utilizing knowledge of the grain structure of wood. Various FEM analysis methods are surveyed, from almost complete reconstruction of the microstructure of wood to much faster, but less precise material parameter approaches.

Computer vision applications are characterized by both high data input rates and high computational costs. Often they need to be realized in a small package, such as a mobile communications device that may consume a limited amount of power for application processing like video coding. This prevents implementations based on conventional processors, even in the future, but hardware acceleration is mandatory. Currently, only monolithic long latency accelerators improve energy efficiency, but they are rigid and costly to design and verify, and difficult to justify for purposes which are considered marginal and computationally expensive.

In 2009, a three year Academy of Finland funded project PARadigm Shift for Embedded Computing (PARSEC) was launched, targeting automation of the tools for designing ultra-energy-efficient systems for computationally demanding purposes, such as video coding and image analysis. The research is done in cooperation with Tampere University of Technology (Prof. Jarmo Takala) and Abo Academi University (Prof. Johan Liljus). We have demonstrated that fine grained, silicon area efficient, adaptable hardware accelerators can be employed at very low software interface overheads through deterministic multithreaded schedules. This has turned out to be much more efficient than the conventional interrupt, semaphore, and event handler mechanisms advocated by the textbooks. In essence, we are targeting a new paradigm for embedded computing and expect significant impacts in the field. Our developments include the simultaneous decoding of multiple MPEG-4 streams on shared accelerators and MIMO reception. The latter work has been carried out in cooperation with the Centre for Wireless Communications (CWC).

Research working towards fine-grained embedded multiprocessing continued during 2009. The previously published approach of efficient multiprocessor scheduling was extended to take into account the memory organization of the underlying platform. Another direction in the development was that of automatic multiprocessor scheduling for Reconfigurable Video Coding decoder descriptions. This approach can also be generalized to other applications that have been described in the CAL actor language.

Magnetic resonance imaging (MRI) of the human body provides 3D-imagery with various contrast options sensitive to different biological phenomena. As such MRI has become an important imaging technology both clinically and for medical sciences. Functional MRI (fMRI) is concerned with indirect imaging of brain activity. fMRI data provides 3D-video on blood oxygenation coupled with electrical activity through metabolic control over vasculature in the brain. Analysis of this 4D-domain combines image analysis with biomedical signal processing.

We have recently started to analyze factors contributing to spatial Independent Component Analysis (sICA) which is currently a widely used analysis method in fMRI. Furthermore, also temporal measures related to complexity and power distribution of signals have been investigated as a means of providing novel clinically useful contrasts based on fMRI data. The research is done jointly with Department of Diagnostic Radiology, Oulu University Hospital.

In addition to several research projects, we have begun developing a Computer Vision Algorithm Library. Over the years, our group has developed numerous algorithms and methods using different development environments and tools. Now these algorithms are being collected into a single C++ library. The intention is to make efficient implementations of the existing algorithms easily reusable for both new research and as part of practical applications like a robot’s vision system. We plan to release this software library in the future with an open source license.

Exploitation of Results

A clear indication of the impact of our research is that many of our papers are frequently cited. According to the ISI Web of Knowledge, the most cited paper from Finland published since 2002 in the ICT-area (electrical & electronic engineering, computer science and telecommunications categories) is our PAMI paper from 2002 (out of about 5500 papers). We have also authored three most cited papers from Oulu in the ICT area published since 1995. Altogether six of our publications are within the top ten, and nine within the top 20 in the same ranking.

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern (LBP) methodology is used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis.

The MVG has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience their roots are especially in the strands of “free academic research”.

There are currently altogether five research based spin-outs founded directly on computer vision area. The number of the spin-outs could be extended up to thirteen when taking into account the influence of the MVG’s almost thirty-year old history and the spin-out companies’ from the spin-out research groups in the area of computer science and engineering in total.

In 2009, one of the most recent spin-outs, Visidon Ltd., released a solution for recognizing objects and faces on mobile phones in real-time. The company also provided face detection technology which supports age, gender and expression recognition for MediaTeam Oulu's UBIOulu project. Together with Symbicon Ltd., Visidon presented a new audience measurement and reporting system, where Visidon’s face technology (age, gender, smile, activity) is integrated into the Symbicon displays and a web-based user interface is provided for easy access.
Future Goals

Research on human-centered vision systems will be further extended. For example, we are launching collaboration with Prof. Tapio Seppänen’s signal analysis team in the area of multimodal analysis of human communicative behavior, combining vision with speech and biosignal data.

The use of multimodal data has also a key role in the new European project called Trusted Biometrics under Spoofing Attacks (Tabula Rasa) in 2010-2014. MVG was invited to be a partner in this project consortium on the basis of the previous contribution the group made for the success of the Mobile Biometry project.

Among our research staff, as much as 40% are currently from abroad. Due to our excellent international reputation, we increasingly attract visitors from abroad to join us for some time, and many of our researchers are willing to make research visits to leading groups abroad. Such bilateral collaboration will bring fresh new ideas and expertise to our research.

Close interaction between basic and applied research has always been a major strength of our group. The scientific output of the group has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.

Personnel

| professors & doctors | 15 |
| graduate students    | 15 |
| others               | 8  |
| total                | 38 |
| person years         | 29 |

External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>514 000</td>
</tr>
<tr>
<td>Ministry of Education</td>
<td>123 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>344 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>238 000</td>
</tr>
<tr>
<td>international</td>
<td>74 000</td>
</tr>
<tr>
<td>total</td>
<td>1 293 000</td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


Chen J, Yi D, Yang J, Zhao G, Li SZ & Pietikäinen M (2009) Learning mappings for face synthesis from near-infrared to visual


Background and Mission

The Machine Vision Group (MVG) is renowned worldwide for its expertise in computer vision. The group has a solid record, which now spans over 29 years, of scientific merits on both basic and applied research on computer vision. The mission of the group is to always pursue actual research challenges and improve the state-of-the-art methods.

MVG works as a single well-focused research group, and intensively collaborates with other groups of similar status in Europe, the USA, China and Japan. Within the Seventh Framework Programme FP7, the group currently participates in a project consortium of Trusted Biometrics under Spoofing Attacks (TABULA RASA). Another project, Mobile Biometry (MOBIO), which was also coordinated by the IDIAP Research Institute, Switzerland, ended in December 2010.

The research areas of the group range from generic computer vision methodologies to machine vision applications and vision systems engineering. The main areas of research are computer vision methods, human-centered vision systems and vision systems engineering. The results of the group’s research have been widely exploited in industry, and contract research forms a part of its activities.

Highlights and Events in 2010

During 2010, the group produced a large number of significant scientific contributions and published its key results through major conferences and journals. The group has established its position among the most significant research teams in the field of machine vision. One of the indicators of this is that many of its papers are frequently cited. Its landmark paper “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, which was published in the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) journal in 2002, has now well over 600 citations according to the ISI Web of Knowledge. It is the most cited Finnish paper published since 2002 in electrical & electronic engineering and all computer science categories, and the 6th most cited paper of the prestigious PAMI journal.

In addition, visibility outside academic forums has been achieved. The group has a tradition of presenting regularly its research projects and activities in different media. For example, a story on the human-robot interaction research, conducted jointly by the Machine Vision Group and the Intelligent Systems Group, appeared on Prisma Studio science program on Finnish TV channel YLE TV1 on March 23, 2010.

The group hosts regularly visits by respected and renowned scientists from abroad. In 2010, the group had the pleasure of hosting Professor Maja Pantic from the Imperial College London, UK, Professor Nicu Sebe from the University of Trento, Italy, Professor Jan Flusser from the Academy of Sciences of the Czech Republic, Dr. Matthew Blashko from the University of Oxford, UK, Dr. Peter Sturm from INRIA Rhône-Alpes, France, and Dr. Christopher Mei from the University of Oxford, UK. In addition, several domestic visitors from both public and private sectors were briefed about our research activities.

The group has established active collaboration with some of the world’s leading institutions and top scientists. It has had collaboration with the University of Maryland (USA) from the early 1980’s. More recent partners include the Chinese Academy of Sciences, the Academy of Sciences of the Czech Republic, the Czech Technical University in Prague, the École Polytechnique Fédérale de Lausanne (EPFL), the University of Oxford, the University of Tokyo and the consortium of the European Mobile Biometry (MOBIO) project. In 2010, MVG established contacts with Eurecom (France), the University of Turin (Italy) and the consortium of the new TABULA RASA project.

The group fosters international mobility to and from our unit. Two of our researchers made 1-3 month research visits to foreign institutions during the reporting period. The group has attracted visiting postdoctoral researchers and postgraduate students from abroad, who are affiliated with us for periods of a couple of weeks up to several years. Two such short visits were made to MVG in 2010 and two long visits began. In addition, MVG recruited two doctoral students from abroad.

The group and its members are active in the scientific community. For example, in 2010 Prof. Pietikäinen was invited to join the editorial board of the Image and Vision Computing journal. Dr. Abdenour Hadid co-organized the first Algerian summer school on biometrics, which was held in Algiers, Algeria on May 23-27, 2010. He also lectured a tutorial on image and video descriptors at the International Conference on Image Processing Theory, Tools and Applications (IPTA 2010) which was held in Paris, on July 7-10, 2010.

The professors of the group were committee members of several major conferences and many researchers of the
group served as reviewers for various journal and conference articles. Dr. Guoying Zhao and Prof. Matti Pietikäinen served as co-editors for a book entitled “Machine Learning for Vision-Based Motion Analysis: Theory and Technologies”, Springer, 2011, together with Dr. Liang Wang and Dr. Li Cheng. Dr. Guoying Zhao was also serving as a guest editor of the special issue New Advances in Video-based Gait Analysis and Applications: Challenges and Solutions on IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, together with Dr. Liang Wang, Dr. Nasir Rajpoot and Prof. Mark Nixon. Dr. Esa Rahtu served as a member of the Board of the Pattern Recognition Society of Finland.

The Pattern Recognition Society of Finland ranked Timo Ahonen’s PhD thesis “Face and texture image analysis with quantized filter response statistics” as the best Finnish PhD thesis made on pattern recognition in 2008-2009. This thesis was also awarded as the best PhD thesis of the University of Oulu in 2009.

Scientific Progress
The current main areas of the research are: 1) Computer vision methods, 2) Human-centered vision systems, and 3) Vision systems engineering.

Computer Vision Methods
The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. In the last few years, the research in computer vision methods has been broadened to cover also two new areas: computational imaging and object detection and recognition. The aim in all these areas is to create a methodological foundation for development of new vision-based technologies and innovations.

Texture Analysis
Texture is an important characteristic of many types of images and can play a key role in a wide variety of applications of computer vision. A textured area in an image can be characterized by a non-uniform or varying spatial distribution of intensity or color. The variation reflects some changes in the scene being imaged. The MVG has long traditions in texture analysis research and ranks among the world leaders in this area. The Local Binary Pattern (LBP) texture operator has been highly successful in numerous applications around the world and has inspired plenty of new research on related methods, such as the Local Phase Quantization (LPQ) and Weber Law Descriptor (WLD), also proposed by researchers from MVG. The writing of a book entitled “Computer Vision Using Local Binary Patterns”, authored by M. Pietikäinen, A. Hadid, G. Zhao and T. Ahonen, was in its final stages in 2010. The book will be published by Springer in 2011.

In the reporting year, a novel learning framework for image descriptor design was proposed. The Fisher separation criterion (FSC) is used to learn the most reliable and robust dominant pattern types considering intra-class similarity and inter-class distance. Image structures are thus described by a new FSC-based learning (FBL) encoding method. LBP was taken as an input to this framework, but it can be easily generalized for other purposes by utilizing different histogram descriptors. FBL-LBP differs from the previous LBP approaches since the FBL framework learns robust dominant types of each class instead of using fixed pattern types. The proposed approach outperformed other comparative methods in the classification experiments on three commonly used texture databases.

We have also continued our research on dynamic texture segmentation and recognition of actions utilizing spatio-temporal texture descriptors. The results of this research will be published in 2011 - 2012.

The FCS-based learning framework.

For video analysis, a novel descriptor called volume local phase quantization (VLPQ) was developed. VLPQ is an extension to the LPQ descriptor used for 2D image analysis, and the method is based on binary encoding of the quantized phase information of the local 3D Fourier transform. The VLPQ descriptor is highly tolerant to image blur in the case of centrally symmetric spatial blurring caused by, for example, misfocused optics or camera motion. The method has been used for dynamic optics classification, and the results achieved are promising. The VLPQ descriptor is not only shown to be blur insensitive, but efficient also in the case of sharp video frames.

Texture is a key feature in the visual diagnosis of variety of medical specimens. Manual inspection of specimens with
a light microscope is still to date the standard procedure. In collaboration with Dr. Johan Lundin’s research group from Institute for Molecular Medicine Finland (FIMM), we have been developing computer-aided diagnosis tools to introduce ways of focusing and speeding-up this quite labor intensive manual workload of pathologists. The LBP texture descriptor was successfully used to automatically classify cancerous tissues into stromal and parenchymal tissue compartments using the Support Vector Machine (SVM) classifier. Other applications include parasite detection and patient outcome prediction based on machine vision tools. In addition to research on analysis algorithms, our collaboration includes developing of a mobile diagnostic device, based on mobile microscopy, which could be used particularly in developing countries.

Computational Imaging

Computational imaging is a relatively new term and it refers to the convergence of the camera and the computer. It provides more flexibility to image acquisition and enables more advanced techniques to be employed that go beyond the capabilities of traditional photography. Images can be processed in many ways, for example, to synthesize new views from a set of input frames. As the actual image is reconstructed in the computer, computational cameras can also utilize unconventional optical solutions for image acquisition. Due to its wide range of applications computational imaging (also referred to as computational photography) has become a major field of research within both the computer vision and the computer graphics communities.

In our research, we have investigated means for utilizing computer vision in various problems of computational imaging. One important problem is transparency, which complicates object segmentation and produces artifacts in free viewpoint video. The current methods for alpha matte estimation rely heavily on user input and smooth image statistics. Following the success of multi-view stereo methods, we developed a multi-view alpha matting method that requires no user input and makes no assumption on image statistics. It analyzes the 3D structure of the scene to find several samples of the same physical point as well as the background color. It is then able to formulate constraints on the components and estimate the true foreground color and level of transparency.

Stereo matching is a fundamental problem in free viewpoint rendering, as well as in computer vision, but it still remains an open problem. Recent patch based multi-view reconstruction approaches are able to accurately reconstruct a well textured scene. However, non-textured surfaces are not reconstructed by such methods, thus resulting in a semi-dense depth map for each image. In order to complement these methods, we have developed a hole-filling algorithm that produces a fully dense depth map from a semi-dense one. It uses planar models to fit a surface to the homogenous regions, taking into account the reconstructed geometry on its boundary.

Knowledge of the scene type can provide important information for image acquisition. In our recent work, we have studied different approaches that can be used in recognizing landscape scenes. The experiments indicate that a global texture-based approach competes with or outperforms other more complex methods in the landscape image recognition problem. Especially the computational cost of the method relying on Local Binary Pattern (LBP) representation is minimal compared to the GIST and bag-of-words methods investigated in the study.

Automatic image quality evaluation can be used to provide immediate feedback about the image quality to the photographer. In our research, we have developed a novel framework for no-reference image quality evaluation. The method uses seven image degradation features that are extracted and fed to a classifier that decides whether the image has good or bad quality based on properties such as blur, under or over exposure, saturation, and lack of meaningful information. The method was designed to be computationally efficient in order to enable real-time performance in embedded devices.

Flash-no-flash imaging is a branch of computational imaging that aims to combine the natural color contents of ambient light images with the details available in all the input images. Aligning images taken with flash and without flash is a challenging problem, where contents of image pairs can differ quite radically. Flash removes natural shadows and creates new ones altering the gradient directions and strengths, and also the colors may change radically. Surfaces may reflect the flash light resulting in anomalies compared to ambient light scenes. Many conventional approaches are unsuitable for reliable flash-no-flash image registration due to these changes.

We have developed a new image aligning method utilizing mutual information driven point matching refinement for the flash-no-flash image pair registration. Initial point matches obtained with block wise phase correlation are refined using iterative location adjustment with increase in mutual information as the driving force. The approach generally outperforms the tested reference methods in terms of registration accuracy for this special case. Visual inspection also confirms that the proposed method succeeds to register flash-no-flash image pairs in cases where other state-of-art approaches fail. We are further interested in the fusion of regular low ambient light images with infrared images taken with and without flash.
We recently considered the problem of visibility improvement in outdoor environments. The main cause of impairments in visibility is often represented by atmospheric scattering, which derives from the presence in the air of aerosol particles, and results in what is commonly perceived as fog, haze, as well as rain or snow. Several current methods try to remove the haze layer from a single image by using a physical model of atmospheric scattering and solving an inverse problem. However, the quality of the de-hazed images strongly depends on the accuracy of the parameters that must be estimated: scene depth, and airlight color. The latter typically gives rise to color artifacts that manifest locally in the processed image. We have developed a novel method to robustly estimate the airlight color from a single image and improve the de-hazing results, eliminating artifacts that are still visible in state-of-the-methods.

Object Detection and Recognition

Object recognition and detection refer to the tasks of classifying and localizing a given object from an image or video sequence. Since the problem in its general form is very difficult to tackle, it is usually constrained to a particular aspect with restricted object class and detection tolerance. For example, one can focus on detecting the bounding boxes containing all persons in the given image. In another example, one could be interested in the pixel level segmentations of cars in aerial images acquired over a city. Our work in MVG mainly concentrates on detecting salient image segments and humans.

Generally, current object detectors scan the image on several scales and examine all sub-windows using a pre-trained class specific classifier. Depending on the classifier, this might be a computationally very expensive approach, and it can still be extended only to a small part of the possible image sub-windows. Furthermore, the process has to be repeated for each object class to be detected. An attractive alternative would be to first use some light processing to identify a set of salient focus points, and then concentrate the computationally intensive methods only to these locations. It is also likely that the human visual system proceeds in this way.

In our research, we have developed methods for locating salient image segments in a computationally efficient manner. The new approach is based on measuring the local feature contrasts in a novel probabilistic framework. Compared to other state of the art saliency detector, our method offers more reliable segmentations which are also in line with the human annotators. Furthermore, the method is also applicable to video sequences, where the issue of computational complexity of class specific detectors is often prohibitively large.

For human detection, we have presented a new descriptor, which is mainly targeted at still images. It is referred to as isotropic granularity-tunable gradients partition (IGGP), which is extended from granularity-tunable gradients partition (GGP) descriptors. The isotropic representation is achieved by aligning the features with different orientation channels according to their principal angles. The benefits of this extension are twofold: firstly, the noise introduced by the small partitions in the original GGP descriptors is eliminated and the performance can be essentially improved; secondly, the integral image based fast computation is applied and more than 20 times speedup has been achieved.

A novel descriptor for human detection in video sequences has been developed. It is referred to as the spatial-temporal granularity-tunable gradients partition (STGGP), which is an extension of granularity-tunable gradients partition (GGP) from the still image domain to the spatial-temporal domain. Specifically, the moving human body is considered as a 3-dimensional entity in the spatial-temporal domain. Then, in 3D Hough space, we define the generalized plane as a primitive to parse the structure of this 3D entity. The advantage of the generalized plane is that it can tolerate imperfect planes with a certain level of uncertainty in rotation and translation. The robustness to the uncertainty is controlled quantitatively by the granularity parameters defined explicitly in the generalized plane.

In the HIMA project that was finished in 2010, a comparative study of recent human detection methods was carried out. The objective was to find the best available human detection methods that could be used in the safety applications of non-road mobile machinery. The study suggested that when considering both accuracy and speed, a good choice would be the well-known HOG based approach, although new approaches such as IGGP can achieve better accuracy, but at the cost of much higher computational complexity.

Human detection methods published in the literature are usually based on analyzing monocular images or image sequences. Stereo images, on the other hand, provide also the depth information that can be used as complementary measurements for human detection. The analysis framework developed in the HIMA project was based on the fusion of several single view human detectors and 3D filtering of the candidate objects. The framework uses the idea of pixel-wise human probabilities, which are obtained by several separate detection trials following binomial distribution. The framework has been extensively evaluated...
with image sequences of authentic work machine environments, and it has proven to be feasible.

**Geometric Computer Vision**

Geometry is an important aspect of computer vision. The laws of geometry and optics describe how the three-dimensional world is imaged on the camera sensor and, hence, an understanding of imaging geometry is important for the development of automatic image analysis methods. The basis for this understanding is the geometry of multiple views, which we have studied from both theoretical and computational viewpoints. In this area, we have a strong research background in certain basic problems, such as geometric camera calibration and image matching, and our research in these areas spans a period of several years. Recently, we have increased the emphasis on research in image-based modeling and rendering. This work directly builds upon our previous expertise. For example, our earlier approach to quasi-dense image matching could be particularly useful in image-based modeling. The motivation for studying multi-view reconstruction and visualization of 3D scenes lies in the increasing number of applications where such techniques are becoming useful and practical.

One of our recent developments in image-based modeling is the extension of the quasi-dense match propagation algorithm to the case of three views. The results of this work show that using the three-view constraint and the best-first propagation strategy together during the correspondence growing improves the accuracy of matching and reduces the occurrence of outliers. In particular, compared with two-view stereo, the new method is more robust for repeating texture. Further, the approach seems particularly promising because it is able to provide reliable and accurate matches with a single growth stage and without additional outlier filtering or iterative refinement for the matched patches, as required in some other patch-based multi-view stereo methods.

In addition, besides studying conventional passive multi-view stereo reconstruction methods, we have recently explored possibilities for utilizing modern active depth cameras in computer vision applications. For example, range sensors that are capable of operating at a video frame rate, such as time-of-flight cameras, could be useful in human-computer interaction.

**Human-Centered Vision Systems**

In future ubiquitous environments, computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems, enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

**Face Recognition and Biometrics**

Automatic face analysis has become a very active topic in computer vision research as it is useful in several applications, like biometric identification, visual surveillance, human-machine interaction, video conferencing and content-based image retrieval. Face analysis is challenging due to the fact that a face is a dynamic and non-rigid object which is difficult to handle. Its appearance varies due to changes in pose, expression, illumination and other factors such as age and make-up. In 2010, our biometric related research focused on novel face descriptors, face analysis using facial dynamics, face recognition from degraded images, face detection and verification in mobile devices, soft biometrics (e.g. gender/age/ethnicity classification), face recognition under spoofing attacks and speaker recognition using visual information.

In the period 2008-2010, MVG participated in the Mobile Biometry (MOBIO) project funded by the European Commission, aiming at researching new mobile services secured by biometric authentication means. The LBP method developed in MVG played an important role in the project. Starting from November 2010, we are continuing the biometry theme within a new FP7 consortium called TABULA RASA (starting in November 2010 and lasting 42 months). TABULA RASA looks at the vulnerabilities of existing biometric systems to spoofing attacks to a wide range of biometrics including face, voice, gait, fingerprints, retina, iris, vein, electro-physiological signals (EEG and ECG) etc. We have started investigating whether skin reflectance properties could be used to detect possible spoofing attacks against face based biometric systems.
The presence of facial occlusions (e.g., due to sunglasses, hats, scarves, beards etc) is unfortunately quite common in real-world applications, especially when individuals are not cooperative with the system such as in video surveillance scenarios. In 2010, we investigated, in cooperation with Eurecom institute (France), new approaches to improve the recognition of faces occluded by sunglasses and scarves. We propose an efficient approach which consists of first detecting the presence of a scarf/sunglasses and then processing the non-occluded facial regions only. The occlusion detection problem is approached using Gabor wavelets, PCA and support vector machines (SVM), while the recognition of the non-occluded facial part is performed using block-based local binary patterns. Experiments on the AR face database showed that the proposed method yields significant performance improvements compared to existing work on recognizing partially occluded and also non-occluded faces.

Blur is often present in real-world images and significantly affects the performance of face recognition systems. Basically, there are two main approaches for face recognition under blur. The first direction consists of deblurring the face images via deconvolution with an aim of recovering sharp images ready for processing using any conventional face recognition algorithm. Another way of handling the recognition of blurred faces consists of deriving blur-invariant features directly from the blurred images without having to recover the sharp images which, anyway, are not necessarily needed for recognition. In 2010, we continued our joint research with the University of Tokyo for improving the recognition of blurred faces and proposed a new approach which inherits and combines the advantages of the two categories of methods. The idea consists of first reducing the amount of blur in the images via deblurring and then extracting blur-tolerant descriptors (using LPQ) for recognition. We assessed our analysis on real blurred face images (FRGC 1.0 database) and also on face images artificially degraded by focus blur (FERET database), demonstrating significant performance enhancement compared to the state-of-the-art.

In 2010, speaker identification using visual information has also been investigated. Visual information from captured video is indeed important for speaker identification under noisy conditions. Combination of LBP dynamic texture and Edge Map structural features were proposed to take both motion and appearance into account, providing the description ability for spatiotemporal development in speech. Spatiotemporal dynamic texture features of local binary patterns extracted from localized mouth regions are used for describing motion information in utterances, which can capture the spatial and temporal transition characteristics. Structural edge map features are extracted from the image frames for representing appearance characteristics. A combination of dynamic texture and structural features takes both motion and appearance together into account, providing the description ability for spatiotemporal development in speech. In the experiments on the BANCA and XM2VTS databases, the proposed method obtained promising recognition results compared to the other features.

Recognition of Facial Expressions and Emotions

A goal of facial expression recognition is to determine the emotional state of the face regardless of its identity. The face can express emotion sooner than people verbalize or even realize their feelings, and research in social psychology has shown that facial expressions form the major modality in human communication. So facial expression is one of the most powerful, natural and immediate means for human beings to communicate their emotions and intentions.

In facial expression recognition, both Region-of-Interest (ROI) features and geometric features are helpful for representation. We have investigated an effective component-based approach to facial expression recognition by combining ROI and geometric features together. Spatio-temporal local texture features were computed from geometric facial areas centered at 38 detected fiducial interest points (determined by ASM) and the generated histograms were concatenated to describe facial expressions. Furthermore, since not all features are important to facial expression recognition, the problem of how to select the most discriminative features for facial components is addressed by using AdaBoost for boosting the strong appearance and motion features.

Moreover, it is widely agreed that emotion is a multimodal procedure, which can be expressed mainly by facial expression, but also by head and body movements, gestures, speech and some physical characteristics like heart rate and blood pressure. Our recent plan is to collaborate with Prof. Tapio Seppänen’s Biomedical Engineering Group to combine signals from different sensors, and develop robust systems for emotion recognition. Investigation into micro-expression analysis and distinguishing spontaneous facial expressions from acted expressions will also be included.

Visual Speech Analysis and Synthesis

It is well known that human speech perception is a multimodal process. Visual observation of the lips, teeth and tongue offers important information about the place of pronunciation articulation. Techniques for interpreting speech using visual information only (namely, lip-reading) have a wide range of applications in the real world. For in-
stance, such a system can be used to understand someone’s speaking in a highly noisy environment (e.g., among a large crowd or in a moving vehicle). It can also be used to improve the quality of the lives of people with hearing impairments. We propose a novel graph embedding method for the problem of automatic lip-reading. To characterize the temporal connections among video frames of the same utterance, a new distance metric is defined on a pair of frames, and graphs are constructed to represent the video dynamics based on the distances between frames. Audio information is used to assist in calculating such distances. For each utterance, a subspace of the visual feature space is learned from a well-defined intrinsic and penalty graph within a graph-embedding framework. Video dynamics are found to be well preserved along some dimensions of the subspace. Discriminatory cues are then decoded from curves of the projected visual features to classify different utterances.

**Overview of the proposed method.**

Video-realistic speech animation plays an important role in the area of affective human-robot interactions. The goal is to synthesize a visually realistic face that can talk just as we do. In this way, it can provide a natural platform for a human user and a robot to communicate with each other. For instance, when there are interactions between human users and machines, instead of using text or audio only to communicate, an animated talking head synchronized with the audio may attract more attention from users and make such applications more engaging. A visually realistic talking head could make users feel comfortable and natural, and hence improve the quality of the human-machine interactions. Moreover, such synthesis techniques may be used to generate visual cues for audio clips so as to help hearing-impaired people understand machine responses better.

The key for synthesizing a talking head is how to realize a visually realistic talking mouth since most of the dynamic shape and texture changes on the face appear in the mouth area. In our work, video synthesis is achieved by first learning generative models from the recorded speech videos and then using the learned models to generate videos for novel utterances. A generative model considers the whole utterance contained in a video as a continuous process and represents it using a set of trigonometric functions embedded within a path graph. To synthesize a video for a novel utterance, the utterance is first compared with the existing ones to find the phoneme combinations that best approximate the utterance, and the image frames are then synthesized using the learned model.

**Tracking and Recognition in Camera Networks**

An ever increasing number of cameras are being utilized in surveillance to monitor activities in locations like airports, hospitals and office buildings. One major issue of multiple camera surveillance is to detect and track moving objects in each view of a camera network and maintain their identities while moving across the views. Our research has focused on the building blocks of a complete camera network tracking system, including object detection, tracking and re-identification nodes, and a communication infrastructure which binds all the nodes together to form a flexible sensor network.

Tracking objects across non-overlapping camera views is one of the main challenges in object tracking. During the year 2010, we developed a real-time system for tracking and re-recognition of moving objects across views, and published the underlying algorithm in the International Conference on Pattern Recognition (ICPR). The algorithm uses adaptive boosting (AdaBoost), binary classification trees and k-means clustering to build classifiers for feature-based classification of object tracking sequences. The features (color, texture, shape, etc.) extracted from the single-view tracks are clustered and then learned to obtain boosted tree classifiers that provide reliable object re-recognition even on previously unseen views of the object.
We also presented a new solution to the problem of person re-identification. Person re-identification means to match observations of the same person across different time and possibly different cameras. The appearance based person re-identification must deal with several challenges such as variations of illumination conditions, poses and occlusions. Our proposed method is inspired from the spirit of self-similarity. Self-similarity is an attractive property in visual recognition. Instead of comparing image descriptors between two images directly, the self-similarity measures how similar they are to a neighborhood of themselves. The self-similarities of image patterns within the image are modeled in two different ways in the proposed Global Color Context (GCC) method. The spatial distributions of self-similarities with regard to color words are combined to characterize the appearance of pedestrians. Very promising results are obtained on the public ETHZ database compared with the state-of-art.

An overview of the proposed method. Local color features are extracted densely and clustered to form a color codebook (a). Then, the assignments of color features to color words in the codebook are explored. Color features from the same visual color word are marked with the same color in (b). For each color word in the codebook, the spatial occurrence distributions of color self-similarities are learned (c).

Affective Human-Robot Interaction

Mobile robots equipped with cameras and other sensors provide an excellent environment for investigating human-machine interfaces, because communication with home-aid robots developed for service tasks in homes, nursing homes, retirement homes and offices, for example, must be easy and natural. An intelligent robot should be able to detect and identify the user in order to personalize its services and guarantee security, it should recognize the user’s emotions to allow for affective interaction, and it should be able to communicate easily with the user and understand given commands by recognizing speech and gestures. The machine must also learn to change its behavior according to the user and his/her emotional state.

Together with the Intelligent Systems Group, we have collaborative research on Affective Human-Robot Interaction (HRI), supported by the Academy of Finland (2009-2012) and the European Regional Development Fund (2010-2012). The research is divided into three parts dealing with machine vision methodology for HRI, robot embodiment and learning in HRI, and experimental validation of affective HRI. The basic research of the MVG in 2010 focused on developing methodologies for a videorealistic talking face to be used in human-robot communication, and for recognizing human gestures.

The experimental system being developed for HRI research consists of 1) a Segway Robotic Mobility Platform (RMP 100) equipped with computers, cameras, microphones, magnetic field sensors, an avatar display etc., 2) a ubiquitous CSE laboratory environment containing a network of cameras and other sensors embedded in the environment. During the reporting year, a large collection of algorithms developed in the MVG were modified for our new Machine vision algorithm library, from which they can be easily used in HRI project demonstrations. The real-time demonstrations made in 2010 include face and people detection and tracking, determining gender, facial expression and identity of the tracked faces, and applying simple gesture recognition for each tracked person.

Camera-Based Interfaces for Mobile Devices

Designing comfortable user interfaces for mobile phones is a challenging problem, given the limited amount of interaction hardware of the device. However, multiple built-in cameras and the small size of handhelds are under-exploited assets for creating novel applications that are ideal for pocket size devices, but may not make much sense with personal or laptop computers. Studies into alternatives for mobile user interaction have, therefore, become a very active research area in recent years.

A key advantage of using cameras as an input modality is that it enables recognizing the 3D context in real-time, and at the same time provides for single-handed operations in which the users’ actions are interpreted without touching the screen or keypad. For example, the user’s position and gaze can be measured, in order to display true 3D objects even on a typical 2D screen.
In our research on camera-based mobile user interfaces, we have developed a prototype for a virtual 3D user interface based on face tracking. The determining of the user’s position and gaze is analyzed in real time, enabling the device to display true three-dimensional objects even on a typical 2D LCD screen. In the developed interface, the user-motion and camera input enable interactivity by providing real-time feedback on the context, while simultaneously guiding the user actions. The solution lies in the extraction of motion and features from sequential video frames, and this can be augmented with the information provided by integrating the data from different sensors such as accelerometers, gyroscopes and magnetometers. The implementations include a parallel pipeline that uses the mobile Graphical Processing Unit (GPU) integrated on the platforms, reducing the computation and power needs of the application.

In the research area of interactive camera-based applications for mobile devices, we have studied algorithms for implementing visual tracking of unknown objects. The user can select the object to be tracked on the touch screen of the phone. To improve the robustness of tracking, online learning and adaptive techniques have been considered for object detection. In the algorithms special attention is paid to their computational efficiency.

On mobile platforms, we have also studied the use of multimodal sensor information for several user interaction scenarios: face detection based backlight and key lock control, face detection based automatic image capturing mode detection, and automatic activation of a “point and find” application. Recently, we have also applied measurement fusion techniques for controlling an interactive panorama image constructor and motion based image browsing.

**Vision Systems Engineering**

Vision systems engineering research provides guidelines for identifying attractive computing approaches, architectures, and algorithms for useful commercial systems. In practice, solutions from low-level image processing to even equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our industrial visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. We have also applied our expertise to applications intended for smart environments and embedded platforms.

Visual inspection research of automated wood strength grading for sawn timber using machine vision has continued. Earlier we have developed a solution that employs real-time feature extraction, classification, and the Finite Element Method (FEM) combined into an adaptive learning scheme. In new developments, one key area was to determine the correlation between features extracted from images and the actual strength of the board in question. Performance of different classifiers was also studied using only knot based features. Good results were achieved for Finnish Pine in which knots are common and are in many cases one of the most important reasons for reduced strength qualities.

We have continued developing our unusual event detection system. The aim is to create cheap systems that are easy to install unnoticeably. To accomplish this, we have been testing the system with very thin compound eye optics. Compound eye optics introduce a new challenging problem for the system, the image quality is considerably lower than with conventional optics. The work in the future will focus on improving the image quality and thorough testing.

The dramatic increase in the number of mobile consumer devices such as digital cameras, mobile phones and laptops has been made possible by the advances in the energy efficiency of electronics. However, new applications are constantly being invented, and old ones are updated, and this has led to an ever-increasing demand for computational power, which results in higher power consumption of devices. As battery technologies advance only very slowly, improving the efficiency of computations has become the only viable alternative: more calculations need to be done with less power.

Our research on energy efficiency concentrates on this problem from the perspective of signal processing, which is required in practically all mobile devices. Improving the energy efficiency of signal processing systems cannot be accomplished solely in software or hardware, but requires consideration of them both, as well as the interface. Besides the work on improving the energy efficiency of signal processing, our group is also interested in hardware-
software development tool chains with a focus on design automation. This work is done in close cooperation with national and international partners.

The work on design automation and multiprocessing for DSP has continued in 2010. The topic of quasi-static scheduling for Reconfigurable Video Coding reached a new milestone as the first implementation of automated quasi-static scheduling was completed, and is to be published at IEEE ICASSP 2011. The work on Transport Triggered Architecture processors has produced, for example, an implementation of a customized, programmable processor for computing Local Binary Patterns.

We also carried out a performance study on a GPU to fulfill the real-time requirements of the LTE MIMO-OFDM detector. The limited size of the fast on-chip memory resources and the required conditional code execution were found to be the limiting factors for the current GPU implementations. An interesting future solution would be a GPU that is designed specifically for baseband solutions.

In the InterSync project, we have studied energy efficient signal processing methods for condition monitoring of steel wire ropes used in cranes. Various filtering methods and algorithms for detecting faults in the magnetic measurement signal were developed. Promising results were achieved using Wavelet and Hilbert transforms to detect momentary abnormalities. In addition, we have implemented methods on the transport triggered architecture (TTA) signal processor in order to achieve efficiency needed by energy autonomous wireless sensor nodes. The work is now progressing towards pilot installations.

**Exploitation of Results**

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern (LBP) methodology is used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis.

The MVG has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience their roots are especially in the strands of “free academic research”.

There are currently five research based spin-outs founded directly on the computer vision area. The number of the spin-outs could be extended up to thirteen when taking into account the influence of the MVG’s thirty-year history and the spin-out companies’ from the spin-out research groups in the area of computer science and engineering.

**Future Goals**

Research on human-centered vision systems and affective computing will be further extended. For example, we have started collaboration with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with speech and biosignal data.

The research on biomedical imaging is also planned to be extended. MVG is a partner in a new Oulu BioImaging OBI network (obi.oulu.fi), which has been accepted as an associated partner to Euro-BioImaging. The mission of Euro-BioImaging is to create a coordinated and harmonized plan for biomedical imaging infrastructure deployment in Europe.

Among our research staff, as many as 40% are currently from abroad. Due to our excellent international reputation, we increasingly attract visitors from abroad to join us for some time, and many of our researchers are willing to make research visits to leading groups abroad. Such bilateral collaboration will bring new fresh ideas and expertise to our research.

Close interaction between basic and applied research has always been a major strength of our group. The scientific output of the group has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.

Prof. Matthew Turk from the University of California, Santa Barbara, has been selected to receive the 2011–2012 Fulbright-Nokia Distinguished Chair in Information and Communications Technologies. With this funding, he will stay two months at MVG. This will offer great opportunities for future collaboration.

<table>
<thead>
<tr>
<th>Personnel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors, doctors</td>
<td>16</td>
</tr>
<tr>
<td>doctoral students</td>
<td>17</td>
</tr>
<tr>
<td>others</td>
<td>16</td>
</tr>
<tr>
<td>total</td>
<td>49</td>
</tr>
<tr>
<td>person years</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>External Funding</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>550 000</td>
</tr>
<tr>
<td>Ministry of Education and Culture</td>
<td>103 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>99 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>123 000</td>
</tr>
<tr>
<td>international</td>
<td>132 000</td>
</tr>
<tr>
<td>total</td>
<td>1 007 000</td>
</tr>
</tbody>
</table>
Doctoral Theses


Selected Publications


Background and Mission

The Center for Machine Vision Research (CMV) is a creative, open and internationally attractive research unit. It is renowned world-wide for its expertise in computer vision.

The center has a solid record, which now spans for 30 years, of scientific merits in both basic and applied research on computer vision. The mission of the center is to develop novel computer vision methods and technologies that create a basis for emerging innovative applications.

In February 2012 the CMV had three professors, 16 senior or postdoctoral researchers, and 25 doctoral students or research assistants. The unit is highly international: about 40% of our researchers (doctors, PhD students) are from abroad, representing ten different nationalities. Over 80% of its publications in 2011 had a non-Finnish author/co-author. CMV has an extensive international collaboration network in Europe, the USA, and China. The mobility of the researchers to leading research groups abroad, and vice versa, is intense. Within the Seventh Framework Programme FP7, the CMV currently participates in the project consortium of Trusted Biometrics under Spoofing Attacks (Tabula Rasa). It also participates in two European COST actions.

The main areas of our research are computer vision methods, human-centered vision systems and vision systems engineering. The results of the research have been widely exploited in industry, and contract research forms a part of our activities.

Highlights and Events in 2011

The year 2011 was the 30th Anniversary of machine vision research at the University of Oulu. The occasion was celebrated in November in a guest seminar, with Prof. Xilin Chen from the Chinese Academy of Sciences as the keynote speaker, followed by an anniversary dinner.

In conjunction with the 30th Anniversary, our re-naming was announced: we are now the Center for Machine Vision Research, abbreviated to CMV, and no longer call ourselves the Machine Vision Group (MVG). The new name describes better the extent of our activities.

As a highlight, in November the founder and director of the Center for Machine Vision Research (CMV), Professor Matti Pietikäinen, was appointed as an IEEE Fellow. He was given recognition for his contributions to texture and facial image analysis for machine vision. An IEEE Fellow is the highest grade of membership and this is recognized by the technical community as a prestigious honor and an important career achievement. The total number selected in any one year is limited to 0.1% of IEEE members who are eligible to vote.

In summer 2011, Springer released the first edition of the book “Computer Vision Using Local Binary Patterns” authored by Professor Matti Pietikäinen, Adjunct Professors Dr. Abdenour Hadid and Dr. Guoying Zhao and CMV alumnus, Dr. Timo Ahonen. The book covers widely LBP methods and their spatial and spatiotemporal extensions, and gives an insight in their various application areas.

During 2011, the researchers of CMV developed the first method ever to automatically detect spontaneous facial micro-expressions. These novel results gained a great deal of attention among the Finnish academia and in the media, e.g. the Finnish television program Prisma Studio and the online news of the leading Finnish newspaper Helsingin Sanomat. Internationally, the publication on micro-expressions presented at ICCV 2011 turned out to be one of the most popular within the interactive poster session. The novel results have been circulated to media abroad as well, especially in the UK.
One of the leading experts in vision-based human-computer interaction, Professor Matthew Turk, from the University of California, Santa Barbara, received the prestigious Fulbright-Nokia Distinguished Chair in Information and Communications Technologies for 2011-2012. This position allows him to stay for two months in Oulu during the summer 2012. The collaboration will focus on vision-based interaction for natural human-computer interfaces in mobile environments. Prof. Turk already visited Oulu briefly this year to give a guest talk.

The CMV had a visible role in the IEEE Conference on Pattern Recognition and Computer Vision (CVPR), held in Colorado Springs, USA, in June. Prof. Pietikäinen and Prof. Janne Heikkilä held a tutorial on image and video description with local binary pattern variants. Prof. Pietikäinen and Dr. Guoying Zhao co-chaired with Matthew Turk, Liang Wang and Li Cheng a workshop on machine learning for vision-based motion analysis (MLVMA). Since the fall of 2011, Pietikäinen and Zhao have been editing, together with the same co-chairs, a Special Issue on Machine Learning in Motion Analysis for Image and Vision Computing journal. Matti Pietikäinen also continued his term as an associate editor of the same journal. In September, Pietikäinen and Heikkilä lectured in a tutorial on image and video analysis with LBP variants at the IEEE International Conference on Image Processing (ICIP) held in Brussels. Dr. Abdenour Hadid lectured a tutorial on face analysis using local binary pattern variants at the IEEE International Conference on Automatic Face and Gesture Recognition (FG), held in Santa Barbara, California in March. The professors and senior researchers of the unit were also committee members of several major conferences, and many of our researchers served as reviewers of various journal and conference articles.

The Center for Machine Vision Research was very successful in getting funding from the Academy of Finland. Dr. Guoying Zhao was nominated as an Academy Research Fellow and was granted additional funding for research costs. Dr. Juho Kannala received Postdoctoral Researcher project funding. Both Professors Matti Pietikäinen and Olli Silvén received Academy project funding for four-year periods.

The CMV regularly hosts visits of renowned scientists from abroad. In 2011, the Center had the pleasure of hosting Prof. Dimitris Metaxas from Rutgers University, USA; Prof. Sebastien Lefevre from the University of South-Britany, France; Prof. Francesco Vaccarino from the Politecnich University of Turin, Italy; Prof. Jan Flusser from the Academy of Sciences of the Czech Republic; and Dr. Victor Lempitsky, affiliated with both the Yandex, Russia and the University of Oxford, UK, in addition to Prof. Chen and Prof. Turk.

The Center fosters international mobility to and from our unit. Five of our researchers made research visits to our partner groups during the reporting period. The CMV has attracted visiting postdoctoral researchers and PhD students from abroad, who are affiliated with us for periods of a couple of weeks up to several years. In 2011, the CMV recruited three new postdoctoral researchers and one doctoral student from abroad.

Professor and Vice Rector (Education) Olli Silvén was invited to become a member of the TTA – the Finnish Academy of Technology from October, 2011. Administratively, it operates within Technology Academy Finland, which awards the Millennium Technology Prize. Acting Professor Jari Hannukela was appointed as Secretary of the IEEE Finland Section in May. Dr. Esa Rahlu served as a board member of the Pattern Recognition Society of Finland.

Scientific Progress

The current main areas of the research are: 1) Computer vision methods, 2) Human-centered vision systems, and 3) Vision systems engineering.

Computer Vision Methods

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. In the last few years, the research on computer vision methods has been broadened to cover a further two new areas: computational photography, and object detection and recognition. The aim in all of these areas is to create a methodological foundation for development of new vision-based technologies and innovations.

Texture Analysis

Texture is an important characteristic of many types of images and can play a key role in a wide variety of applications of computer vision. The CMV has long traditions in texture analysis research and ranks among the world leaders in this area. The Local Binary Pattern (LBP) texture operator has been highly successful in numerous applications around the world and has inspired plenty of new research on related methods, including the blur-insensitive Local Phase Quantization (LPQ) method, also developed at CMV.

Recently, we proposed a novel approach to computing rotation invariant features from histograms of local, non-invariant patterns. We applied this approach to both static and dynamic Local Binary Pattern descriptors. For static texture description, we presented Local Binary Pattern Histogram Fourier features (LBP-HF), and for dynamic texture recognition, two rotation invariant descriptors computed from the LBP-TOP (Local Binary Patterns from Three Orthogonal Planes) features in the spatiotemporal domain. LBP-HF is a novel rotation invariant image descriptor computed from discrete Fourier transforms of local binary pattern (LBP) histograms. The approach can also be generalized to embed any uniform features into this framework, and combining supplementary information, e.g. sign and magnitude components of LBP together can improve the description ability. Moreover, two variants of rotation invariant descriptors were proposed for the LBP-TOP, which is an effective descriptor for dynamic texture recognition, but it is not rotation invariant. In the
experiments, it was shown that LBP-HF and its extensions outperform non-invariant and earlier versions of rotation invariant LBP in rotation invariant texture classification. They are also robust with respect to changes in viewpoint, outperforming recent methods proposed for view-invariant recognition of dynamic textures.

They are also robust with respect to changes in viewpoint, outperforming recent methods proposed for view-invariant recognition of dynamic textures.

We also investigated rotation invariant image description with a linear model based descriptor named MiC, which is suited to modeling microscopic configuration of images. To explore multi-channel discriminative information on both the microscopic configuration and local structures, the feature extraction process is formulated as an unsupervised framework. It consists of: 1) the configuration model to encode image microscopic configuration; and 2) local patterns to describe local structural information. In this way, images are represented by a novel feature: local configuration pattern (LCP). The performance of this method was evaluated on textures present in three challenging texture databases: Outex_TC_00012, KTH-TIPS2 and Columbia-Utrecht (CUReT). The encouraging results showed that LCPs are highly discriminative.

Texture is a key feature in the visual diagnosis of medical settings. Manual inspection of specimens with a light microscope is still to date the gold standard. Within the Institute of Molecular Medicine in Finland (through distance work from the University of Oulu), we are exploring with Adjunct Professor Johan Lundin high throughput computer assisted methods for automated analysis of digitized breast and colorectal cancer samples and microbiological samples e.g. malaria parasites.

The LBP texture descriptor was successfully used to classify colorectal samples into stromal and cancerous compartments. The Support Vector Machine (SVM) was used for classification. The automatic segmentation of tissue samples provides a feasible way to compare biomarker expressions in different tissue types.

The morphology of breast cancer tumor tissue is undisputedly related to the outcome of breast cancer. We have developed a method for estimating breast cancer morphological properties from digitized microscopic images using LPQ and LBP texture features and a supervised SVM classifier. We have also developed a method for clustering breast cancer tissue images of different patients based on texture features and unsupervised clustering. It was possible to identify clusters with remarkably high or low average survival of patients, and this can be used to support the diagnosis.

**Computational Photography**

Computer vision as a research area has expanded and evolved during the past decades, and the boundaries with many other disciplines have become blurry. Computer graphics is one of those fields which is closely related to computer vision. In particular, computational photography is a widely studied topic where both communities share a common interest. In computational photography, the aim is to develop techniques for computational cameras that give more flexibility to image acquisition and enable more advanced features to be employed that go beyond the capabilities of traditional photography.

One fresh approach for computational photography is light field imaging. Conventional cameras capture a single intensity image of a view, while light fields store also information on the direction of incoming light rays. Light fields can be used to calculate new representations of the imaged scenes where results are refocused on new focal planes or have changes in view point.

Our goal is to develop new light field imaging devices in collaboration with VTT and PrintoCent. Imaging properties of thermoplastic lenses are studied and compared with
commercially available glass lenses. Light field rendering methods are also studied and developed. Initial results show that polymeric micro lenses can be used for imaging, yet the fabrication process requires some optimization. The implemented rendering methods confirm that single snapshots can be refocused after capturing the image data.

Left: raw lens array image of roundworm transverse section through mid-body. Right: reconstructed worm image.

When capturing images of real-world environments, there are often many causes that can contribute to an impairment of visibility. One type of degradation is caused by unfavorable atmospheric conditions. The presence in the air of aerosols and water droplets decreases the visibility range due to multiple scattering of light, resulting in what we commonly refer to as fog, or haze. Single-image de-hazing methods attempt to recover the original radiance at each pixel, removing the effect of haze. However, they require an accurate estimation of the brightness and color of the air light, in order to produce realistic and artifact-free visual results. We developed a de-hazing method that is based on novel statistics obtained from natural images, and it works reliably within a broad class of images and with less strict assumptions on their content.

Haze removal: from left to right, original hazy image, result of de-hazing with one of the existing methods in the literature, and our method.

Object Detection and Recognition

Humans can effortlessly recognize thousands of object classes, which is crucial for successful interpretation of visual content. Recent advances in computer vision have made automatic object detection more practical, and nowadays it is possible to automatically retrieve images which contain a particular object instance or objects from a certain object class. While impressive results have been demonstrated in some object detection problems, there remain several essential open questions. One of the important issues is related to the scalability of current systems. Modern methods can recognize only a couple of dozen object classes, which is very little compared to human perception. The main sources of these limitations are the computational bottlenecks.

Example results of automatic object detection.

In our recent work, we have also developed a method to estimate saliency in images simulating saccadic eye movements. The main contributions of this research are 1) simultaneous simulation of saccadic eye movements and eye fixation prediction, 2) application of stochastic filtering to bottom-up image saliency. Initially, we developed a method to estimate visual stimuli at a given pixel based in our former research. Thereafter, we developed a system of saccade and fixation estimation based on biological evidence that proves the role of eye movements in salience perception. The method proposed incorporates Bayesian filtering techniques to provide a mechanism for imitating saccadic eye movements. Subsequently, we select fixation points based on the amount of visual stimuli perceived for each saccade. The method was evaluated by using several criteria and it was shown to outperform most of the state-of-the-art saliency methods.

Saliency detection: from left to right, original image, density map from human fixation points, and estimated saliency map.
**Geometric Computer Vision**

Geometric computer vision studies geometric aspects of image formation. Knowledge of geometry is often important for the development of automatic image analysis methods. In particular, such applications that require the computer to observe and interact with its three-dimensional environment benefit from geometric techniques. For example, automatic construction of 3D scene models from multiple photographs is a classical, but still relevant research problem. Further, new active depth cameras, such as the Kinect sensor, have boosted rapid progress in scene modeling, especially in indoor environments where textureless surfaces have traditionally been a problem for passive sensing techniques.

Our group has a strong research background in geometric computer vision. Recently, we have increased our research efforts on two fronts, namely, in conventional multi-view stereo methods and in new active range sensing methods. The former research direction builds upon our previous work on quasi-dense image matching and aims to develop a generic multi-view stereo reconstruction approach. In our recent experiments, we performed a preliminary comparison with the current state-of-the-art and observed that our method produces reconstructions of comparable quality, but substantially faster.

The latter research direction aims to utilize modern active depth cameras, such as Kinect, for 3D modeling of indoor environments. The Kinect device has attracted a great deal of attention in the research community. It captures color and depth simultaneously, which makes it suitable for many applications ranging from 3D scene reconstruction to gaming. However, its proprietary nature limits its flexibility for the research community. For example, the calibration procedure is proprietary, and unavailable to the general public. Aiming to contribute to the research community and explore the full potential of the Kinect device, we developed an algorithm that simultaneously calibrates the depth and color cameras. For this calibration, we developed a novel distortion correction algorithm that achieves more accurate results than the manufacturer. The result of this work has been published as an open source toolbox for the research community. The algorithm was published in June 2011 and it has generated much interest in the Kinect community since then. In fact, the first version of the toolbox has been downloaded about 3000 times during the first ten months following its publication. In future, our plan is to continue studies with active range sensing devices in order to create techniques for improved 3D modeling of environments and for better human-computer interaction in such environments.

**Human-Centered Vision Systems**

In future ubiquitous environments computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

**Face Recognition and Biometrics**

In 2011, we continued our investigations on demographic classification from face videos using manifold learning and obtained very good results. Research on automatic demographic classification is still in its infancy despite the vast potential applications. The few existing works are only based on static images, while nowadays input data in many real-world applications consist of video sequences. From these observations, and also inspired by studies in neuroscience emphasizing manifold ways of visual perception, we proposed a novel approach to demographic classification from video sequences which encodes and exploits the correlation between the face images through manifold learning. Our extensive experiments on the gender and age classification problems show that the proposed manifold learning based approach yields excellent results, outperforming those of traditional static image based methods. Furthermore, to gain insight into the proposed approach, we also investigate an LBP (local binary patterns) based spatiotemporal method as a baseline system for combining spatial and temporal information to demographic classification from videos.

Starting from November 2010, CMV is participating in an FP7 EU project TABULA RASA looking at the vulnerabilities of existing biometric systems to spoofing attacks to a wide range of biometrics including face, voice, gait, fingerprints, retina, iris, vein, electro-physiological signals (EEG and ECG) etc. CMV is playing a key role in the project, and is leading a work package on the evaluation of the vulnerabilities of existing biometric systems when confronted by spoofing attacks. CMV has also contributed to the definition of the specifications of the spoofing databases that are recorded. Research on countermeasures to face and gait spoofing attacks has also continued.

Without anti-spoofing measures most of the state-of-the-art facial biometric systems are vulnerable to attacks, since they try to maximize the discrimination between identities, instead of determining whether the presented trait originates from a real live client. Even a simple photograph of...
the enrolled person’s face, displayed as a hard-copy or on a screen, will fool the system. As an initial counter measure, we proposed to approach the problem of spoofing attacks from a texture analysis point of view, since face prints usually contain printing quality and other recapturing defects, e.g. blur, that can be detected using texture features. Our LBP based method showed excellent preliminary results on several spoofing databases. Furthermore, we took part in IJCB 2011 Competition on Counter Measures to 2D Facial Spoofing Attacks and were able to achieve perfect discrimination between spoofing attacks and real client access.

**Recognition of Facial Expressions and Emotions**

Facial expression recognition is used to determine the emotional state of the face, regardless of its identity. Most of the existing datasets for facial expressions are captured in a visible light spectrum. However, the visible light (VIS) can change with time and location, causing significant variations in appearance and texture. We have done novel research on a dynamic facial expression recognition, using near-infrared (NIR) video sequences and LBP-TOP (Local binary patterns from three orthogonal planes) feature descriptors. NIR imaging combined with LBP-TOP features provides an illumination invariant description of face video sequences. Appearance and motion features in slices are used for expression classification, and for this, discriminative weights are learned from training examples. Furthermore, component-based facial features are presented to combine geometric and appearance information, providing an effective way of representing the facial expressions. Experimental results of facial expression recognition using a novel Oulu-CASIA NIR&VIS facial expression database, a support vector machine and sparse representation classifiers show good and robust results against illumination variations. This provides a baseline for future research on NIR-based facial expression recognition.

We also proposed a weighted component-based feature descriptor for expression recognition in video sequences. Firstly, texture features and structural shape features are extracted in three facial regions: the mouth, cheeks and eyes of each face image. Then, these extracted feature sets are combined using confidence level strategy. A method for automatically learning different weights to components via multiple kernel learning is proposed for taking into account the different contributions of facial components to expression recognition. Experimental results on the Extended Cohn-Kanade database show that our approach, combining a component-based spatiotemporal features descriptor and a weight learning strategy achieves better recognition performance than the state of the art methods.

**Recognition of Facial Expressions and Emotions**

Facial expression recognition is used to determine the emotional state of the face, regardless of its identity. Most of the existing datasets for facial expressions are captured in a visible light spectrum. However, the visible light (VIS) can change with time and location, causing significant variations in appearance and texture. We have done novel research on a dynamic facial expression recognition, using near-infrared (NIR) video sequences and LBP-TOP (Local binary patterns from three orthogonal planes) feature descriptors. NIR imaging combined with LBP-TOP features provides an illumination invariant description of face video sequences. Appearance and motion features in slices are used for expression classification, and for this, discriminative weights are learned from training examples. Furthermore, component-based facial features are presented to combine geometric and appearance information, providing an effective way of representing the facial expressions. Experimental results of facial expression recognition using a novel Oulu-CASIA NIR&VIS facial expression database, a support vector machine and sparse representation classifiers show good and robust results against illumination variations. This provides a baseline for future research on NIR-based facial expression recognition.

Facial micro-expressions are rapid involuntary facial expressions which reveal suppressed affect. We proposed the first framework to successfully recognize spontaneous facial micro-expressions. For this research, we also designed an induced emotion suppression experiment to collect a spontaneous micro-expression corpus (SMIC). Inside the framework, we use the temporal interpolation method (TIM) to counter short video lengths, LBP-TOP to handle dynamic features and \{SVM, MKL, RF\} to perform classification. We are now expanding the SMIC corpus to more participants.

**Recognition of Facial Expressions and Emotions**

Facial expression recognition is used to determine the emotional state of the face, regardless of its identity. Most of the existing datasets for facial expressions are captured in a visible light spectrum. However, the visible light (VIS) can change with time and location, causing significant variations in appearance and texture. We have done novel research on a dynamic facial expression recognition, using near-infrared (NIR) video sequences and LBP-TOP (Local binary patterns from three orthogonal planes) feature descriptors. NIR imaging combined with LBP-TOP features provides an illumination invariant description of face video sequences. Appearance and motion features in slices are used for expression classification, and for this, discriminative weights are learned from training examples. Furthermore, component-based facial features are presented to combine geometric and appearance information, providing an effective way of representing the facial expressions. Experimental results of facial expression recognition using a novel Oulu-CASIA NIR&VIS facial expression database, a support vector machine and sparse representation classifiers show good and robust results against illumination variations. This provides a baseline for future research on NIR-based facial expression recognition.

Facial micro-expressions are rapid involuntary facial expressions which reveal suppressed affect. We proposed the first framework to successfully recognize spontaneous facial micro-expressions. For this research, we also designed an induced emotion suppression experiment to collect a spontaneous micro-expression corpus (SMIC). Inside the framework, we use the temporal interpolation method (TIM) to counter short video lengths, LBP-TOP to handle dynamic features and \{SVM, MKL, RF\} to perform classification. We are now expanding the SMIC corpus to more participants.

**Recognition of Facial Expressions and Emotions**

Facial expression recognition is used to determine the emotional state of the face, regardless of its identity. Most of the existing datasets for facial expressions are captured in a visible light spectrum. However, the visible light (VIS) can change with time and location, causing significant variations in appearance and texture. We have done novel research on a dynamic facial expression recognition, using near-infrared (NIR) video sequences and LBP-TOP (Local binary patterns from three orthogonal planes) feature descriptors. NIR imaging combined with LBP-TOP features provides an illumination invariant description of face video sequences. Appearance and motion features in slices are used for expression classification, and for this, discriminative weights are learned from training examples. Furthermore, component-based facial features are presented to combine geometric and appearance information, providing an effective way of representing the facial expressions. Experimental results of facial expression recognition using a novel Oulu-CASIA NIR&VIS facial expression database, a support vector machine and sparse representation classifiers show good and robust results against illumination variations. This provides a baseline for future research on NIR-based facial expression recognition.

Facial micro-expressions are rapid involuntary facial expressions which reveal suppressed affect. We proposed the first framework to successfully recognize spontaneous facial micro-expressions. For this research, we also designed an induced emotion suppression experiment to collect a spontaneous micro-expression corpus (SMIC). Inside the framework, we use the temporal interpolation method (TIM) to counter short video lengths, LBP-TOP to handle dynamic features and \{SVM, MKL, RF\} to perform classification. We are now expanding the SMIC corpus to more participants.
We also proposed a method to successfully differentiate spontaneous versus posed (SVP) facial expressions with a realistic corpus, achieving very promising results. Our method uses graph embedding to temporally interpolate image sequences, and inputs the resulting frames through a new spatiotemporal local texture descriptor CLBP-TOP into a set of classifiers. We showed that the CLBP-TOP outperforms other descriptors, and that SVP differentiation benefits from both temporal interpolation and near-infrared images. We further propose a new generic framework for facial expression recognition for solving the general facial expression analysis problem.

Visual Speech Analysis and Synthesis
Video-realistic speech animation plays an important role in the area of affective human-robot interactions. The goal of such animation technology is to synthesize a visually realistic face that can talk just as we do. In this way, it can provide a natural platform for a human user and a robot to communicate with each other. In addition, the techniques also have potential applications such as generating synchronized visual cues for audios in order to help hearing-impaired people better capture information, or to make human characters in movies.

It is known that human speech perception is a multi-modal process which makes use of information not only from what we hear (acoustic) but from what we see (visual). In machine vision, visual speech recognition (VSR) is the task of recognizing the words uttered by a speaker through analyzing a video of the speaker’s talking mouth without audio input. VSR is an important alternative to traditional speech recognition technology in human-machine interactions, especially when audio is unavailable or severely corrupted by noise.

We consider the same practical VSR problem of classifying words, phrases or short sentences. The solution to the problem can be widely used in applications such as a social robot or a car stereo control system, to facilitate human computer interactions. We use a path graph to represent the sequential structure of a visual speech signal. A novel model is constructed to connect the extracted visual features with a low-dimensional continuous curve embedded in the graph.

We first record a video corpus within which a human character is asked to speak different utterances. The mouth is then cropped from the original speech videos and is used to learn generative models for synthesizing novel mouth images. A generative model considers the whole utterance contained in a video as a continuous process and represents it using a set of trigonometric functions embedded within a path graph. The transformation that projects the values of the functions to the image space is found through graph embedding. Such a model allows us to synthesize mouth images at arbitrary positions in the utterance. To synthesize a video for a novel utterance, the utterance is first compared with the existing ones from which we find the phoneme combinations that best approximate the utterance. New mouth images are then synthesized from the learned models, based on the combinations. Finally, we project seamlessly the synthesized mouth back to some background video to gain realism.

Recognition of Actions
We presented a new method for robust recognition of complex human actions. We first cluster each video in the training set into temporal semantic segments by a dense descriptor. Each segment in the training set is represented by a concatenated histogram of sparse and dense descriptors. These histograms of segments are used to train a classifier. In the recognition stage, a query video is also divided into temporal semantic segments by clustering. Each segment will obtain a confidence evaluated by the trained classifier. Combining the confidence of each segment, we classify this query video. To evaluate our approach, we performed experiments on two challenging datasets, i.e., the Olympic Sports Dataset (OSD) and the Hollywood Human Action dataset (HOHA). We also tested our method on the benchmark KTH human action dataset. Experimental results confirmed that our algorithm performs better than the state-of-the-art methods.

Some example frames of human action videos from OSD (the top two rows) and HOHA (the bottom two rows).
Affective Human-Robot Interaction

Research on affective human-robot interaction (HRI) is made with the support of the Academy of Finland (2009-2012) and the European Regional Development Fund (2010-2012), in collaboration with the Intelligent Systems Group. An experimental HRI platform working in a smart environment has been developed, including a Segway Robotic Mobility Platform (RMP 200), equipped with laptop computers, Kinect depth sensors and video cameras, microphones, magnetic field sensors, an avatar display, and a ubiquitous multi-camera environment with wireless access to the Internet. The development of the platform was continued in 2011. Computer vision methods for different tasks were developed, including methods for localization, obstacle detection, facial image analysis, and human-robot interaction.

A real-time version of our visual speech synthesis algorithm was implemented that is capable of generating video-realistic speech animation from text input at interactive speeds. Language support was also extended from English only to Finnish using the original training material. This version of synthesis was integrated in the experimental HRI platform to be used as an integral part of the robot avatar. The Avatar is representing the robot as a human face that communicates with the user with synthesized audio and corresponding synchronized mouth movements.

Our mobile robot will operate in a smart laboratory environment containing a distributed network of cameras. Object re-recognition in camera networks is a major challenge in computer vision. A comprehensive amount of data is required for thorough algorithm development and testing. In 2011, a large real-life database was collected using the laboratory’s indoor IP camera network. The database consist of 100 objects (persons) appearing in 1-5 different camera views, thus making the total number of videos over 400. The labeled collection of surveillance videos provides a practical platform for developing and comparing object tracking and re-recognition methods.

Camera-Based Interfaces for Mobile Devices

Improving usability and user experience with mobile phones is a challenging problem, given the limited amount of interaction hardware of the device. However, multiple built-in cameras and the small size of handhelds are under-exploited assets for creating novel applications that are ideal for pocket size devices, but may not make much sense with personal or laptop computers. Studies into alternatives for mobile user interaction have, therefore, become a very active research area in recent years. A key advantage of using cameras as an input modality is that it enables recognizing the 3D context in real-time, and at the same time provides for single-handed operations in which the users’ actions are interpreted without touching the screen or keypad. For example, the user’s position and gaze can be measured, in order to display true 3D objects even on a typical 2D screen.

In our research on camera-based mobile user interfaces, we have constructed a mobile application prototype where the determination of the user’s position and gaze is analyzed in real-time, a technique that enables the display of true 3D objects even on a typical 2D LCD screen. In the developed interface, we have integrated a series of interaction methods where the user motion and camera input realistically control the viewpoint on a 3D scene. The head movement and gaze can be used to interact with hidden objects in a natural manner just by looking at them. The solution lies on the extraction of features from sequential video frames and the complementary information provided by integrating the data from different motion sensors. The implementation includes a parallel pipeline that reduces the latencies and power needs of the application by using the mobile Graphics Processing Unit (GPU) integrated on the platforms.
On mobile platforms, we have studied the use of multimodal sensor information for several user interaction scenarios such as face recognition based backlight and key lock control. From the beginning of 2011, we have been collaborating with the Nokia Research Center in a project that aims to develop a user interface for mobile devices that utilizes gesture recognition. The gestures are recognized from the front camera and touch screen. With the user interface, the user can move the mouse cursor, click on objects and scroll documents. The functions provided to the user depend on the distance between the hand and the device. The gestures that require more accuracy are detected from the touch screen and gestures that do not need the user to select any specific object on the screen, such as scrolling the screen, are detected from the front camera.

**Vision Systems Engineering**

Vision systems engineering research provides guidelines to identify attractive computing approaches, architectures, and algorithms for useful commercial systems. In practice, solutions from low-level image processing to equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our industrial visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial systems. We have also applied our expertise to applications intended for smart environments and embedded platforms.

Visual inspection research of automated wood strength grading for sawn timber using machine vision has continued. Earlier we have developed a solution that employs real-time feature extraction, classification, and a Finite Element Method (FEM) combined into an adaptive learning scheme. In new developments, the key area has been to determine correlation between features extracted from images and actual strength of the board in question. Performance of the classification was improved by combining knot and grain features. Good results were achieved for Finnish Pine in which knots are common and are in many cases one of the most important reasons for reduced strength qualities.

Sawmill industry related machine vision applications have a long history in our research. One of the latest solutions is the lumber tracing system using cameras installed into the sawmill line. The developed method applies CS-LBP feature vectors formed using gradient images taken from boards before and after the drying process. Tests using image data from several hundreds of actual boards collected from Finnish sawmills have yielded excellent results.

The dramatic increase in the number of mobile consumer devices such as digital cameras, mobile phones and laptops has been made possible by the advances in electronics energy efficiency. However, new applications are constantly being invented, and old ones are being updated, which has led to an ever-increasing demand for computational power. This in turn results in higher power consumption of devices. As battery technologies advance only very slowly, improving the efficiency of computations has become the only viable alternative: more calculations need to be done with less power.

The Academy of Finland awarded a 4-year research grant for a consortium project named DORADO. The project creates tools for generating efficient embedded software/hardware solutions. Platform independent high-level specifications are used to describe parallelism at different levels (data, instruction, task and memory levels). The target is many-core systems that are becoming the key approach in improving the computing throughput. The automation of the hardware design process is emphasized, ultimately for the generation of efficient multi-core application-specific processors. The expected results are high impact techniques for designing and programming heterogeneous systems: automated, platform-independent development tool chains that exhibit “performance portability” across different computing platforms and platform variations. Furthermore, the research is expected to produce analysis methods and tools to automatically interpret the behavior of applications and to streamline their performance, including scheduling of memory accesses at multi-core and single core levels.

Work on design automation and multiprocessing for DSP has continued in the Energy Efficient Architectures and Signal Processing research branch. A demonstration of automatic synthesis of multiprocessor systems to FPGA platforms was completed in early 2011 and published subsequently in the IEEE SiPS Workshop. The work was carried out together with INSA Rennes (France). Further results in cooperation with French researchers are ex-
pected in the near future. Research on automated scheduling of applications written in the RVC-CAL dataflow language has also continued in cooperation with INSA Rennes. Practical cooperation in this direction will also be carried out with the Embedded Systems Laboratory of Åbo Akademi. Research on Transport Triggered Architecture (TTA) processors produced an application-specific instruction processor targeted for extracting Local Binary Pattern features. Ongoing work includes the development of a fully programmable ZigBee baseband processor in the TTA technology.

Our research on energy efficiency concentrates on this problem from the aspect of signal processing, which is required practically in all mobile devices. Improving the energy efficiency of signal processing systems cannot be done solely in software or hardware, but requires consideration of both, as well as of the interface. Besides the work of improving the energy efficiency of signal processing, our unit is also interested in hardware-software development tool chains with a focus on design automation. In the InterSync project, a rapidly reconfigurable energy efficient wireless sensor node implementation on a low-power Flash FPGA was realized. In addition, the energy efficiency of the sensor node designs between fixed versus floating point arithmetic was evaluated. The wireless sensor node was implemented based on transport triggered architecture, and very low energy consumption was achieved. The wireless sensor node was designed for rolling bearing condition monitoring but since it is designed for general purpose signal processing, it could be used also in various other applications.

**Exploitation of Results**

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern (LBP) methodology is used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis. The researchers in CMV have actively published the source codes of their algorithms to the research community and this has increased the exploitation of our results. For example, in 2011 we released a Matlab toolbox for geometric calibration of Kinect with an external camera that has received much interest from the other researchers worldwide.

The results have also been utilized in our own projects. For example, we have started collaboration with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with physiological biosignals. Together with Prof. Seppänen and Dr. Seppo Laukka (Department of Educational Sciences and Teacher Education) and Prof. Matti Lehtihalme (Faculty of Humanities) we are also participating in the FSR Second Wave project where we have developed a Mobile Multimodal Recording System (MORE) that will be used in classroom research in various schools.

CMV is a partner in a new Oulu BioImaging OBI network (obi.oulu.fi), and this network has been accepted as an associate partner in Euro-BioImaging, which aims at creating a coordinated and harmonized plan for the deployment of biomedical imaging infrastructure in Europe. In collaboration with Biocenter Oulu, we have started a new project where the aim is to apply computer vision methods in various problems of biomedical image analysis and to develop a service that provides new image analysis tools for researchers.

Most of our funding for both basic and applied research comes from public sources such as the Academy of Finland and Tekes, but besides these sources, CMV also conducts research by contract that is funded by companies. In this way, our expertise is being utilized by industry for commercial purposes, and even in consumer products, like mobile devices.

The CMV has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are spin-out enterprises. According to our experience, their roots are especially in the strands of “free academic research”. There are currently altogether five research based spin-outs founded directly on the computer vision area. The number of the spin-outs could be extended up to thirteen when taking into account the influence of the CMV’s thirty-year history and the spin-out companies’ from the spin-out research groups in the area of computer science and engineering.

**Future Goals**

Among our research staff, already about 40% are from abroad. Due to our excellent international reputation, we increasingly attract visitors from abroad to join us for some time, and many of our researchers are willing to make research visits to leading groups abroad. Such bilateral collaboration will bring fresh new ideas and expertise to our research.

The two-month visit of the Fulbright-Nokia Distinguished Chair, Professor Matthew Turk, in summer 2012 is expected to bring new ideas into our research on vision-based interaction for natural human-computer interfaces in mobile environments. This visit will provide a basis for further collaboration with the University of California.

We have also other plans to strengthen our research on multimodal human-computer interaction. New proposals on related topics have been submitted to the EU and Tekes. We are also participating in new European project proposals in biometrics.

Close interaction between basic and applied research has always been a major strength of our research unit. The scientific output of the CMV has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.
Selected Publications


Background and Mission

The Center for Machine Vision Research (CMV) is a creative, open and internationally attractive research unit. It is renowned world-wide for its expertise in computer vision.

The center has a strong record, which now spans for over 31 years, of scientific merits in both basic and applied research on computer vision. It has achieved ground-breaking research results in many areas of its activity, including texture analysis, facial image analysis, geometric computer vision, and energy-efficient architectures for embedded systems. The mission of the center is to develop novel computer vision methods and technologies that create a basis for emerging innovative applications.

In February 2013 the CMV had three professors, one FiDiPro professor, 14 senior or postdoctoral researchers, and 26 doctoral students or research assistants. The unit is highly international: about 43% of our researchers (doctoral and post-doctoral) are from abroad. CMV has an extensive international collaboration network in Europe, the USA, and China. The mobility of the researchers to leading research groups abroad, and vice versa, is intense. Within the Seventh Framework Programme FP7, the CMV currently participates in the project consortium of Trusted Biometrics under Spoofing Attacks (Tabula Rasa). It also participates in two European COST actions.

The main areas of our research are computer vision methods, human-centered vision systems and vision systems engineering. The results of the research have been widely exploited in industry, and contract research forms a part of our activities.

Highlights and Events in 2012

The Center for Machine Vision Research (CMV) witnessed its most successful year ever in high impact journal acceptances in the year 2012. Altogether 16 journal articles were accepted for publication. Seven were co-authored articles with our international colleagues, and the remaining nine were written by our research staff. Out of the overall amount, as many as ten were published in highly prestigious journals.

CMV’s paper “Description of interest regions with local binary patterns” with Marko Heikkilä as the first author, and Prof. Matti Pietikäinen and Dr. Cordelia Schmid (INRIA Grenoble) as the co-authors, was given the Best Paper Award among all papers published in 2009 in the prestigious Pattern Recognition journal. The award ceremony was held at ICPR 2012, Tsukuba, Japan.

Professor Xilin Chen from the Chinese Academy of Sciences (CAS) joined CMV as FiDiPro Professor in August 2012. He will spend several short periods of time in Oulu during the FiDiPro (the Finland Distinguished Professor Programme) funding period of August 2012 – February 2016. The “Perceptual interfaces for intelligent human-computer interaction” project will combine the expertise of the world-renowned computer vision research groups in Finland and China.

The CMV Leader, Professor Matti Pietikäinen received The Pentti Kaitera Prize 2012 for his outstanding achievements in machine vision research as well as for his significant contribution to advancing welfare in Northern Finland.

As a new application area, CMV’s computer vision methods are now applied in biomedical science. The Finnish Funding Agency for Technology and Innovation (Tekes) granted two-year funding for an interdisciplinary project between Biocenter Oulu’s Tissue Imaging Center and CMV. The objective is to develop algorithm-based video image analysis tools and machine learning techniques that combine multidimensional image information with the different type of biomedical data.

In summer 2012, CMV had the pleasure of hosting in Oulu a two-month visit of one of the leading experts in vision-based human-computer interaction, Professor Matthew Turk from the University of California, Santa Barbara. Prof. Turk held the prestigious Fulbright-Nokia Distinguished Chair position in Information and Communications Technologies in 2011–2012.
As an essential part of the visit, Prof. Turk gave a course on mobile computer vision within the Infotech Oulu Doctoral Program. In addition, similar PhD level courses were given by Adjunct Professor, Dr. Kari Pulli, from NVIDIA Research, and Dr. L’ubor Ladicky, from the University of Oxford, UK.

The CMV co-organized an International Workshop on Computer Vision with Local Binary Pattern Variants (LBP 2012) in conjunction with the Asian Conference on Computer Vision (ACCV 2012), in early November in Daegon, Korea. The workshop provided the state of the art and the most recent developments in the use of LBPs and their variants in computer vision. The CMV Leader, Professor Matti Pietikäinen, held the keynote at the venue, as well as one of the three keynotes within the International Conference on Image and Signal Processing (ICISP 2012), in Agadir, Morocco, in late June.

CMV hosted the Technical Meeting of a European FP7 project Trusted Biometrics under Spoofing Attacks (shortly Tabula Rasa) in June. This was the first time the project partners had gathered in Oulu.

The developments in affective human robot interaction research raised wide interest in the Finnish media in mid-August, when professors Matti Pietikäinen and Juha Rönöng (ISG) showcased the social robot Minotaurus and its capabilities in the Academy of Finland’s media event. Numerous newspapers and the main tech-releases covered the Minotaurus robot understanding speech commands and responding with an avatar appearing on its display.

Scientific Progress

The current main areas of research are: 1) Computer vision methods, 2) Human-centered vision systems, and 3) Vision systems engineering. In addition to these main areas, our research activities have been recently extended to biomedical image analysis where we collaborate with Biocenter Oulu.

Computer vision methods

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. In the last few years, the research in computer vision methods has been broadened to cover also two new areas: computational photography and object detection and recognition. The aim in all these areas is to create a methodological foundation for development of new vision-based technologies and innovations.

Texture analysis

Texture is an important characteristic of many types of images and can play a key role in a wide variety of applications of computer vision and image analysis. The CMV has long traditions in texture analysis research and ranks among the world leaders in this area. The Local Binary Pattern (LBP) texture operator has been highly successful in numerous applications around the world, and has inspired plenty of new research on related methods, including the blur-insensitive Local Phase Quantization (LPQ) method, also developed at CMV.

Effective characterization of texture images is always an important issue which requires the exploitation of spatial correlations between pixels. As two commonly used texture descriptors, Local Binary Patterns (LBPs) reflect the co-occurrence of binary comparisons among pixels within a local area, whereas Covariance Matrices (CovMs) statistically capture correlation among elementary features of pixels over a certain image region. Enhanced performance is expected if these two kinds of information could be combined in a compact descriptor. Unfortunately, though CovMs are capable of blending multiple informative features in compact and powerful descriptors, the discriminative LBP features could not be exploited as elementary features for CovMs directly, since ordinary LBPs are not numerical variables in Euclidean spaces. Hence local co-occurrence, brought by LBP-like features, and global correlation, brought by CovMs, could not be combined to achieve enhanced discriminative power. To address this problem, we developed a powerful descriptor, named COV-LBP. Firstly, a variant of LBPs in Euclidean spaces, named the LBP Difference feature (LBPD), which can be used for any statistical image region description was proposed. LBPD reflects how far one LBP lies from the LBP mean of a given image region. It is simple, descriptive, rotation invariant, and computationally efficient. Secondly, by applying LBPD in multiple commonly used elementary features mapped from the original image, we provided a bank of discriminative features optional for CovMs. Consequently, the information of LBPs and CovMs are embedded in a unified COV-LBP descriptor. The performance of COV-LBP was evaluated on textures present on three challenging texture databases: Outex_TC_00012, KTH-TIPS, and KTH-TIPS2a. The encouraging results showed that COV-LBP provides high discrimination and good robustness.

Different features extracted from the “Lena” image: (a) the original image; (b) the ordinary LBP feature image in the intensity channel; and (c) the signed LBP Difference image. LBPDs are complementary to the original elementary features.
As one of the major problems in computer vision, texture classification has shown significant improvements; however, the extraction of effective features for texture image representation is still considered as a challenging problem. To obtain discriminative patterns, we presented a learning model which is formulated into a three-layered model. It estimates the optimal pattern subset of interest by simultaneously considering the robustness, discriminative power and representation capability of features. This model is generalized and can be integrated with existing LBP variants such as conventional LBP, rotation invariant patterns, local patterns with anisotropic structure, completed local binary pattern (CLBP) and local ternary pattern (LTP) to derive new image features for texture classification. The derived descriptors were extensively evaluated on two publicly available texture databases (Outex and CUReT) for texture classification, two medical image databases (Hela and Pap-smear) for protein cellular classification and disease classification, and a neonatal facial expression database (infant COPE database) for facial expression classification. Experimental results demonstrate that the obtained descriptors led to state-of-the-art classification performance.

Dynamic texture (DT) is an extension of texture to the temporal domain. How to segment DTs into disjoint regions is a very challenging problem. DTs might be different from their spatial mode (i.e., appearance) and/or temporal mode (i.e., motion field). To this end, we developed a framework based on the appearance and motion modes. For the appearance mode, we use a new local spatial texture descriptor to describe the spatial mode of DT; for the motion mode, we use the optical flow and the local temporal texture descriptor to represent the temporal variations of DT. In addition, for the optical flow, we use the Histogram of Oriented Optical Flow (HOOF) to organize them. To compute the distance between two HOOFs, we developed a simple, effective and efficient distance measure based on Weber Law. Furthermore, we also addressed the problem of threshold selection by proposing a method for determining thresholds for the segmentation method by offline supervised statistical learning. Experimental results show that our method provides very good segmentation results compared to the state-of-the-art methods in segmenting regions that differ in their dynamics. The results were recently published in IEEE Transactions on Image Processing. Following this work, we further studied how to improve the performance and efficiency of our approach. This was achieved by computing the histogram of the spatiotemporal local texture descriptor in one volume and employing the segmentation results of the previous frame for the segmentation of the current frame. The results were published at the ICPR 2012 conference.

We also continued our research on dynamic texture synthesis. A new approach was proposed which is centered on synthesis in the spatial domain, unlike previous work that is mostly focused on the temporal domain. The method explores a 3D patch based synthesis, where the patch selection is accomplished by using spatiotemporal LBP-TOP features, instead of just using the intensity of pixels. The experiments showed that our approach provides a good description of the structure and motion of dynamic textures without generating visible discontinuities or artifacts.

In texture recognition, classes of textures are commonly obtained by creating databases of samples captured under controlled conditions; however when a given image has to be classified, there might be no guarantee of its visual quality and no knowledge of the process of acquisition; thus, such an image might have undergone several types of degradations, for example geometric or radiometric distortions, noise, and blur. We studied the problem of color texture classification under blur and illumination changes. We extended our local-phase-quantization (LPQ) descriptors to generate blur-robust features of color images. The extension is derived from a representation based on Clifford algebra that treats color as a quaternionic quantity. The main advantage of this representation is that it easily
allows generating descriptors that are invariant to illumination changes and yet still preserve the blur-robustness of the grayscale LPQ descriptors. The proposed color texture descriptor achieves superior accuracy over its grayscale counterpart and other color texture descriptors like LBP. Furthermore, we observed remarkable performances in challenging scenarios of varying illumination, without the need to preprocess textures with color-constancy algorithms.

Computational photography

In computational photography, the aim is to develop techniques for computational cameras that give more flexibility to image acquisition and enable more advanced features to be employed, going beyond the capabilities of traditional photography. In 2012, we have studied three topics related to computational photography: microlens array imaging, blur invariant registration, and image priors.

In microlens array imaging, we have introduced a new imaging setup, where the microlenses capture directly the microscopic targets without additional refractive optics. The setup utilizes a single aperture between the sample and the illumination source so that no micro apertures are needed for individual micro lenses. We have shown that this type of setup reaches 8.8 µm resolution with hot embossed microlenses, and 5 µm with glass lenslets, and this resolution is already enough for detecting Schistosoma parasite eggs in urine samples.

The light field rendering methods are studied further, and more emphasis is put on obtaining accurate calibration of the individual lenses. Good calibration should enable better rendering quality, while also various super resolution approaches are under the loupe.

Prior models are useful in many image processing applications (e.g. denoising, inpainting, stereo, optical flow, etc.) They encapsulate our knowledge and assumptions about the structure of images. Image priors can be applied to any modality (e.g. color, infrared, depth). Each modality, however, has its own statistics and thus follows a different model. In the context of 3D reconstruction, we often have a color image with an aligned semi-dense depth map. We developed a joint prior for both intensity and depth, thus taking advantage of the implicit relations between the two channels. This prior can be used in many applications. For instance, we have applied it to the problem of depth map inpainting and upsampling.

Object detection and recognition

Humans can effortlessly recognize thousands of object classes, which is crucial for successful interpretation of visual content. Recent advances in computer vision have made automatic object detection more practical, and nowadays it is possible to automatically retrieve images which contain a particular object instance or objects from a certain object class. Our research on this subject in 2012 was concentrated on two important subtopics, namely fine grained object classification and salient object localization.

Conventional object classification considers classes that are clearly distinct from each other, e.g. cars, bicycles, cats, persons. Another, almost equally important task is to
distinguish between fine-grained sub-classes of the same parent class. We call the resulting classification problem a fine-grained object classification. A good example is flower classification, where the task is to tell the exact species of the flower in the image. Another useful example is bird classification, where we try to identify the exact bird species. The both of these problems are difficult, since there exists a vast quantity of different flower and bird species in the world, and often there are only subtle differences between classes.

Compared to conventional object classification, the fine-grained version has some important characteristics. Usually the appearance difference that distinguishes one class from the other is small compared to the natural variations caused by other factors like natural variations and deformation within the class, background, lighting conditions, viewpoint, and scale. Because of these, it is important for a successful algorithm to be able to eliminate (in one way or another) as many of these variations as possible.

One of the most successful strategies in fine-grained classification has proven to be the following: 1) learn an accurate high level segmentation algorithm for the parent class, and 2) then learn a highly discriminative classifier that operates only with the segmented objects. Although, it might sound straightforward, it turns out that it is extremely difficult to develop accurate segmentation methods for the parent class. Moreover, the required training data (e.g. segmented flower images) is usually costly to obtain.

To this end, we have developed, together with the Visual Geometry Group at the University of Oxford, an automatic co-segmentation approach that is able to learn the necessary parameters using only weakly labeled images; namely images for which we know only the class label (no segmentation boundaries or bounding boxes etc.). This kind of data is rather cheap to obtain using for example a Google image search. Moreover, our approach promotes areas which are important in distinguishing between the fine-grained classes. In our work, we were able to show a considerable improvement with several fine-grained image classification datasets.

During the last year, we exploited the ideas from neural adaptation in the human perception system in order to measure motion saliency in videos. Decorrelation techniques let us discard redundant data in a small temporal window. This will produce the areas of the image that are prone to motion. The system was prototyped in Matlab and tested on the Background Model Challenge (BMC) data set. The method ranked 3rd in the BMC competition held in conjunction with the 11th Asian Conference on Computer Vision (ACCV 2012).

We also exploited saliency for object tracking. To this end, we defined a local similarity number operator which measures the amount of saliency of a pixel in terms of the number of similar pixels in the surrounding area. Afterwards, a target patch is modeled using a joint Saliency-Color distribution and tracked. The proposed method outperforms similar techniques where the target patch is presented by textural-color and color distribution.

Examples of different target representation methods (from top to bottom: color patch, texture-color, and saliency color). As depicted, saliency-color patches are more uniform and focus more on the details of the specific target.

Geometric computer vision

Creating intelligent machines which are capable of complex information processing tasks is one of the great challenges of computer science. Thanks to the development of science, such machines are gradually coming closer to reality. However, in order to become reality, smart machines need better methods for observing their environment and interacting with it, and developing such methods requires knowledge of computer vision and geometry. Our research in geometric computer vision supports the goal of smart machines, and allows improvement in the ways that computers are able to interact with humans and the environment. Hence, geometric computer vision can be seen as a complementary field to our research in human-centered vision systems.

Our group has an extensive research background in various problems of geometric computer vision and we have continued our research efforts on several fronts. For example, we have recently developed a new multi-view stereo reconstruction method which acquires three-dimensional point cloud reconstructions of scenes from multiple photographs. This method was published in ICPR 2012, and it utilizes a prioritized matching strategy in which the most promising point correspondences between the different views are established first, and then they are iteratively expanded in the best-first order. To the best of our knowledge,
our approach is the first one that uses prioritized matching and is able to directly utilize all the available input views. Comparison to the state of the art shows that our method produces point cloud reconstructions of comparable quality, but substantially faster. Hence, it is a useful tool for creating accurate image-based 3D models of scenes.

An example of a point cloud generated from three input views using our method. Prioritized matching expands the correct matches and produces an accurate point cloud (bottom). The lines represent a sparse set of correct (green) and incorrect (red) correspondences.

In addition to multi-view stereo, which refers to passive reconstruction methods solely based on conventional photographs, we have studied active depth sensing techniques by using a Kinect device that utilizes structured infrared light. Such active depth sensing techniques are particularly important in indoor environments where large texture-less surfaces typically cause problems for purely image-based reconstruction methods. As a first step, we have developed an approach to accurate geometric calibration of Kinect devices. Our approach includes a depth-dependent distortion model, and it allows more accurate modeling of the device than competing approaches. The proposed calibration method was recently published in the IEEE TPAMI journal, and the related software is available as an open-source Matlab toolbox. This toolbox has already created widespread interest in the research community, and its different versions have been downloaded more than 5,000 times.

When using active depth cameras, like the Kinect, for acquiring three-dimensional models of indoor environments, it is often necessary to combine long sequences of overlapping depth maps. A simple merging of the points results in much redundant data, which slows down further processing and requires more resources. We have recently worked on a method for creating a non-redundant point cloud with varying levels of detail. The method does not limit the captured volume or require any parameters from the user. Furthermore, overlapping measurements are used to refine the point estimates so that the accuracy of the resulting non-redundant point cloud is better than that of the individual input depth maps. The developed method helps us to produce accurate and compact models of indoor environments more efficiently.

We have also studied techniques for image-based camera localization in indoor environments. In our recent study, we focused on the indoor image matching problem in which the scene information is gathered from multi-modal (2D/3D) sensors. Today, extracting such information from real environments is rather effortless with the help of the new generation depth cameras and range scanners such as Kinect. In the literature, indoor scene matching is considered to be more challenging than the outdoor scene matching problem because outdoor scenes contain more discriminative and unique features leading to comparatively easy recognition. Besides, indoor scenes comprise many similar structures such as doors, windows, chairs, etc. Therefore we proposed combining local 2D and 3D descriptors and explored different ways of combining them. The local properties of 3D data are less distinguishing than their texture-based counterparts. To overcome this limitation we employed glocal (Global-Local) 3D descriptors. The proposed method achieved better performance than the state-of-the-art methods employing only 2D data.

In future, we plan to continue our efforts in image-based modeling, as our recent results and the current state of the field provide a good basis for impressive future applications. For example, we aim to combine conventional multi-view stereo and modern active depth sensing techniques in order to provide better and more convenient tools for building image-based 3D models, and interacting with them.

Human-centered vision systems

In future ubiquitous environments, computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems, enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

Face recognition and biometrics

CMV has continued to play a key role in the FP7 EU project Tabula Rasa (started in 2010 and ending in 2014), looking at the vulnerabilities of existing biometric recognition systems to spoofing attacks on a wide range of biometrics,
including face, voice, gait, fingerprints, retina, iris, vein, electro-physiological signals (EEG and ECG) etc. CMV was the leader of the work package (WP) on the evaluation of biometric systems under spoofing attacks, and also contributed to the development of countermeasures to face and gait spoofing attacks. A joint paper with Prof. Mark Nixon’s group from the University of Southampton, discussing whether gait biometrics can be spoofed or not, was accepted for oral presentation at the International Conference on Pattern Recognition (ICPR 2012).

Without anti-spoofing measures, most of the state-of-the-art facial biometric systems are indeed vulnerable to attacks, since they try to maximize the discrimination between identities, instead of determining whether the presented trait originates from a real live client. Even a mere photograph of the enrolled person’s face, displayed as a hard-copy or on a screen, will fool the system. We have been approaching the problem of spoofing attacks from a texture analysis point of view, since fake faces usually contain recapturing defects, e.g. blur and spoofing medium artifacts that can be detected using texture features. As an initial countermeasure, we proposed using fusion of LBP and gray-level co-occurrence matrices (GLCM) based features, because the combined description provided an effective representation of the overall facial texture quality. Furthermore, we extended our micro-texture analysis based spoofing detection into the spatiotemporal domain and introduced a dynamic texture based face liveness description consisting of both facial appearance and dynamics. More specifically, local binary patterns from three orthogonal planes (LBP-TOP) were utilized for describing specific dynamic events, e.g. facial motion patterns and sudden characteristic reflections of planar spoofing media, and scenic cues which might differentiate real faces from fake ones. Since motion is an important visual cue in spoofing detection, a significant performance enhancement was obtained when the facial dynamics information was exploited in addition to facial appearance.

In 2012, we also continued our research on demographic classification with emphasis on novel local binary pattern variants, especially for gender and age classification. Among the significant achievements is a simple yet efficient extension to LBP which gives a significant improvement compared to the conventional LBP method. Our extensive experiments showed very promising results, not only in gender and age classification, but also in other problems such as texture classification and face recognition. The proposed extension is based on denser image sampling with respect to the LBP neighborhood’s center reference. The method turned out to perform well with other LBP variants, for example, with CLBP and LTP which are among the most powerful LBP variants.

CMV has co-organized a very successful international workshop on computer vision with local binary patterns variants was held in Daejeon, South Korea, on November 5th, 2012.

Recognition of facial expressions and emotions
Facial expression recognition is used to determine the emotional state of the face, regardless of its identity. Feature representation is an important research topic with facial expression recognition in video sequences. We proposed to use spatiotemporal monogenic binary patterns to describe the appearance and motion information of the dynamic sequences. Firstly, we use monogenic signals analysis to extract the magnitude, the real picture and the imaginary picture of the orientation of each frame, since the magnitude can provide much appearance information, and the orientation can provide complementary information. Secondly, the phase-quadrant encoding method and the local bit exclusive operator are utilized to encode the real and imaginary pictures from orientation in three orthogonal planes, and the local binary pattern operator is used to capture the texture and motion information from the magnitude through three orthogonal planes. Finally, both the concatenation method and multiple kernel learning method are exploited to handle the feature fusion. The experimental results on the Extended Cohn-Kanade and Oulu-CASIA facial expression databases demonstrate that the proposed methods perform better than the state-of-the-art methods, and are robust to illumination variations.

Visualization of the original image, the local monogenic magnitude binary pattern, the local monogenic real image binary pattern, the local monogenic imaginary image binary pattern.
In addition, for the dynamic facial expression recognition problem, we proposed a new scheme formulating the dynamic facial expression recognition problem as a longitudinal atlases construction and a deformable groupwise image registration problem. Longitudinal atlases of each facial expression are constructed by performing groupwise registration among all the expression image sequences of different subjects. The constructed atlases can reflect overall facial feature changes of each expression among the population, and can suppress the bias due to inter-personal variations. This method was extensively evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS dynamic facial expression databases. Experimental results demonstrate that our method consistently achieves the highest recognition accuracies among other methods under comparison on all the databases.

Facial occlusion is a challenging research topic in facial expression recognition (FER). It leads us to develop some effective facial representations and occlusion detection methods in order to extend FER to uncontrolled environments. It should be noted that most of the previous work is focused on these two issues separately, and on static images. We were thus motivated to propose a complete system consisting of facial representations, occlusion detection, and multiple feature fusion in video sequences. For achieving a robust facial representation due to the contributions of facial components to expressions, we proposed an approach deriving six feature vectors from eyes, nose and mouth components to form a facial representation. These features with temporal cues are generated by dynamic texture and structural shape feature descriptors. On the other hand, occlusion detection is realized by the traditional classifiers or model comparison. Recently, sparse representation has been proposed as an efficient method of combating occlusion, while it is correlated with facial identity in FER, unless an appropriate facial representation is being used. Thus, we presented an evaluation that demonstrates that the proposed facial representation is independent of facial identity. We then exploited sparse representation and residual statistics to occlusion detection of the image sequences. As concatenating six feature vectors into one causes the curse of dimensionality, we proposed multiple feature fusion, consisting of a fusion module and weight learning. Experimental results on the Extended Cohn-Kanade database and simulated database demonstrate that our framework outperforms the state-of-the-art methods for FER in normal videos, and especially, in partial occlusion videos.

We also continued our research on micro-expression analysis. Micro expressions are short, involuntary facial expressions which reveal hidden emotions. They are important for understanding humans’ deceitful behavior. Currently attention is elevated in both academic fields and in the media. However, while general facial expression recognition (FER) has been intensively studied for years in computer vision, little research has been done in automatically analyzing facial micro-expressions. The biggest obstacle to date has been the lack of a suitable database. We built a novel Spontaneous Micro-expression Database SMIC, which includes 164 micro-expression video clips elicited from 16 participants. Micro-expression detection and recognition tests were carried out by using LBP-TOP as the feature descriptor and the SVM as the classifier and test performance are provided as baselines. SMIC provides sufficient source material for comprehensive testing of automatic systems for analyzing micro-expressions; this has not been possible with any previously published database. The SMIC database is now available from the CMV webpage (http://www.cse.oulu.fi/SMICDatabase).
We also proposed a new method for encoding local binary patterns using a re-parametrization (RP) of the second local order Gaussian Jet. The information provided by RP generates robust and reliable histograms, and is thus suitable for different facial analysis tasks. The proposed method has two main processes: the RP process, which is used to compute needed parameters in a video sequence, and the encoding process, which combines the textural information provided by the LBP and the robustness of the re-parametrization. We showed that this approach can be used for recognizing facial micro-expressions from videos, obtaining competitive performance on the Spontaneous Micro-expression Corpus (SMIC) and the YORK Deception Detection Test.

![A block diagram summarizing the different steps for computing the encoding using re-parametrization.](image)

A project dealing with multimodal emotion recognition for affective computing, funded by the National Agency for Technology and Innovation (Tekes), was in progress. This is a joint effort with Prof. Tapio Seppänen’s Biosignals team, aiming at fusing facial expression information with physiological signals. Some promising preliminary results with realistic data were presented in the EMBC 2012 conference.

**Visual speech animation**

Video-realistic speech animation plays an important role in the area of affective human-computer/robot interactions. The goal of such animation technology is to synthesize a visually realistic face that can talk just like as we do. In this way, it can provide a natural platform for a human user and a robot to communicate with each other. Besides that, the techniques also have potential applications, such as generating synchronized visual cues for audios in order to help hearing-impaired people better capture information, or making human characters in movies.

For this research, we first recorded a video corpus within which a human character is asked to make different utterances. His/her mouth is then cropped from the original speech videos and used to learn generative models for synthesizing novel mouth images. A generative model considers the whole utterance contained in a video as a continuous process and represents it using a set of trigonometric functions embedded within a path graph. The transformation that projects the values of the functions to the image space is found through graph embedding. Such a model allows us to synthesize mouth images at arbitrary positions in the utterance. To synthesize a video for a novel utterance, the utterance is first compared with the existing ones from which we find the phoneme combinations that best approximate the utterance. When selecting video segments for synthesis, we loosen the traditional requirement of using triphone as the unit to allow segments to contain longer natural talking motion. Dense videos are sampled from the segments, concatenated and downsampled to train a video model which enables efficient time-alignment and motion smoothing for the final video synthesis. Different viseme definitions are used to investigate the impact of visemes on the video realism of the animated talking face.

Facial expression is one of the most cogent, naturally preeminent means for humans to communicate emotions, to clarify and stress what is said, to signal comprehension, disagreement, and intentions. Human-machine interaction can benefit significantly from utilizing an “emotional” information channel in the form of facial expressions in addition to speech. In order to extend visual speech synthesis to emotional speech, a new corpus of video data was recorded where a human subject makes different utterances in a certain emotional state. Then different areas of the face from the recorded video can be taken into account in the animation stage to generate images for the novel video, preserving emotions of the face that are realistic and dynamic. The mouth area still remains the most important as it contains visual information both from expressions and speech.

**Human tracking and action recognition**

Tracking objects in a camera network requires a large set of testing data and methods for algorithm validation. In 2012, we published a dataset called CMV100, and a few AdaBoost-based baseline methods for object re-identification in camera networks. The dataset contains 100 tracked objects and more than 400 videos. It consists of the original surveillance videos, foreground masks and an extensive amount of tracking data. Various image descriptors (color, texture, shape, etc.) are also provided for each object. The dataset is publicly available at our website (http://www.ee.oulu.fi/research/imag/cmv100/), along with software tools for processing the tracking data.

![Sample images from the CMV100 dataset.](image)

As a part of the Future School Research Second Wave project, we have been developing a mobile multimodal recording system, called MORE. The system consists of several microphones and cameras that can be used to record events in different learning environments. The acquired material is synchronized and can be easily browsed and analyzed afterwards to support pedagogic and didactic purposes.
Affective human-robot interaction

Research on affective human-robot interaction (HRI) has been made with the support of the Ubiquitous computing and diversity of communication (MOTIVE) program of the Academy of Finland (2009–2012) and the European Regional Development Fund (2010–2013), in collaboration with the Intelligent Systems Group. An experimental HRI platform, called Minotaurus, working in a smart environment has been developed, including a Segway Robotic Mobility Platform (RMP 200), equipped with laptop computers, Kinect depth sensors and video cameras, microphones, magnetic field sensors, an avatar display, and a ubiquitous multi-camera environment.

The development of the platform was further continued in 2012 and a robotic arm was integrated into the platform. Computer vision methods for different tasks were developed and integrated, including methods for localization, obstacle detection, facial image analysis, audio-visual speech synthesis, and human-robot interaction. One of the highlights in 2012 was the wide press coverage in national print and online media, following the “science breakfast” organized by the Academy of Finland for media representatives.

In 2009, we started developing a machine vision algorithm library on top of the popular OpenCV library. Currently it contains a wide variety of different algorithms, and has been successfully used as a part of the robot’s vision system. We have also developed a generic configurable processing node for processing image sequences from different image sources, including networked security cameras and Microsoft’s Kinect sensors attached to the moving robot. The algorithm set, image sources and all the relevant tuning parameters are dynamically configurable over the network. This enables us to adapt to changing environment and processing needs. After processing images, a node provides information about the found objects and their properties to the other parts of the system. This information can then be used, for example, to navigate the robot towards humans, avoid obstacles and visualize the environment for users. Our camera network itself has been extended with new cameras and all the cameras are now calibrated. This enables us to estimate the 3D coordinates of all detected objects moving in the environment.

The first demonstrations on the whole Minotaurus system, operating in a smart environment, were presented in December 2012.

Camera-based interfaces for mobile devices

Improving usability and user experience with handheld mobile devices is a challenging problem, given the limited amount of interaction hardware of the device. However, multiple built-in cameras and their small size are under-exploited assets for creating novel solutions that are ideal for pocket size devices, but may not make much sense with desktop computers. Studies into alternatives to mobile user interaction have, therefore, become a very active research area in recent years. A key advantage of using cameras as an input modality is that it enables recognition of the 3D context in real-time, and at the same time provides for single-handed operations in which the users’ actions are interpreted without touching the screen or keypad. For example, the user’s position and gaze can be measured, in order to display true 3D objects even on a typical 2D screen.

In the research area of interactive mobile applications, we have continued the research on multimodal gesture controlled user interaction. The user interface works with the already existing hardware in recent mobile devices. The gestures are recognized from the front camera and the touch screen. With the user interface, the user can move the mouse cursor, click on objects and scroll documents. The functions provided to the user depend on the distance between the hand and the device. For this purpose, we have developed a new finger detection and tracking system based on color features.

We have also studied implementation of motion-based segmentation. The approach is based on estimating the displacement of a set of feature points. The algorithm developed exploits the similarity of consecutive video frames and uses estimates of feature displacements to propagate segmentation information from frame to frame. Performance of the segmentation is improved by exploiting information about the uncertainty of displacement estimates. The method was utilized in our hand gesture controlled user interface.
On mobile platforms, we have investigated the combination of interactive imaging and energy-efficient high performance computing to enable new user interactions. Using cameras as an input modality provides single-handed operations in which the users’ actions are recognized without interactions with the screen or keypad. The solution analyses and compares the means to reach interactivity and performance with sensor fusion and asymmetric parallel processing, taking advantage of the multiple computing resources present on the current mobile platforms such as Graphics Processing Units and Digital Signal Processors. We have constructed an application prototype where the determination of the user’s position and gaze is analyzed in real time, a technique that enables the display of true 3D objects even on a typical 2D screen. In the developed interface, we have integrated a series of interaction methods where the user motion and camera input realistically control the viewpoint on a 3D scene. The head movement and gaze can be used to interact with hidden objects in a natural manner just by looking at them.

Camera based strength grading research was also continued. Earlier we have developed a solution that employs real-time feature extraction, classification, and the Finite Element Method (FEM) combined into an adaptive learning scheme. Now, methods for grain edge detection in low quality images under challenging lighting conditions have been explored. New ways to use grain based information in visual strength grading were also tested. The ultimate goal is to find the answer of whether camera based strength grading can achieve the same accuracy as conventional mechanic methods like bending machines. Good results were achieved for Finnish Pine, in which knots are common, and are in many cases one of the most important reasons for reduced strength qualities.

The work on the energy efficient architectures and signal processing topic has been carried out, strongly supported by the DORADO project funded by the Academy of Finland. The project creates tools for generating efficient embedded software/hardware solutions. Platform independent high-level specifications are used to describe parallelism at data, instruction, task and memory levels. The target is many-core systems that are becoming the key approach in improving computing throughput. The automation of the hardware design process is emphasized, ultimately for the generation of efficient many-core application-specific processors. The expected results are high impact techniques for designing and programming heterogeneous systems: automated, platform-independent development tool chains that exhibit “performance portability” across different computing platforms and platform variations.
Energy-efficient, and yet programmable, solutions for various applications have been developed. The H.265 HEVC video coding methodology that is going to be standardized in January 2013 contains a computationally very challenging part called the Adaptive Loop Filter (ALF). In our group, ALF was implemented with a programmable processor that is capable of real-time processing at HDTV resolutions. Similarly, a programmable, yet efficient solution for a ZigBee baseband radio was finished. The programmable ZigBee radio enables placing the communication infrastructure on the same programmable chip together with other applications; this opens up new possibilities for creating wireless sensor nodes for ubiquitous computing. Finally, a programmable, but efficient solution for real-time extraction of Local Binary Patterns was finalized in our group.

Our research on energy efficient signal processing for wireless sensor nodes has also continued. The research has been concentrating on node designs, but the same principles can be applied also in general purpose designs. A real-life demonstration together with VTT was carried out by deploying our signal processing module on a complete Flash FPGA based sensor node solution by VTT. The I/O interfaces for our transport triggered architecture (TTA) based signal processing module were implemented, and optimized designs for most common signal processing tasks such as FFT were developed and implemented using TTA.

Multiprocessing design automation was continued, together with the French INSA research institute in Rennes. INSA has adopted the work started in our group and has been continuing the development of multiprocessors for video processing, both independently and together with our group. In our group, this work has resulted in an automated tool chain that maps a program to several customized signal processors that can be placed on an FPGA board. Starting from 2013, the research topic of energy efficient architectures and signal processing will be strengthened through the US-Finnish cooperative project CREAM that combines energy-efficient computations with dataflow-based design automation.

**Biomedical image analysis**

Analysis of medical and biological images is an important application area for computer vision. We have recently started collaboration with Biocenter Oulu where the aim is to apply state-of-the-art computer vision algorithms to research problems in cell biology. Modern bioimaging results in an enormous amount of data, and efficient extraction of available information using existing computational image analysis tools has emerged as a significant bottleneck. In a joint project funded by Biocenter Finland and the University of Oulu, we have started to set up and develop a novel customizable image analysis service that could be provided to researchers working with biological images.

In late 2012, we also started a new project called “Algorithm-based combination and analysis of multidimensional video and open data” (ABCdata), funded by Tekes (strategic research opening). In this project, the objective is to analyze 3D microscopic image sequences, and develop tools for cell segmentation and tracking, as well as for detection of cellular events such as mitosis and apoptosis in conditions that mimic human tissues, which makes this research unique from the scientific point of view.

Cell migration analysis is an essential tool when making comparisons between the effects of different drugs for the treatment of diseases. Traditionally, the tracking task has been carried out by manually annotating the cells in time-lapse microscopy images. It is easy to understand that this kind of approach is very laborious and error prone when handling large numbers of cells over long time periods. Phase-contrast microscopy is the most commonly used imaging technique to visualize living cells, due to the simple configuration of microscope instruments, low costs, and cell visualization without the use of fluorescent labels and phototoxicity. The aforementioned imaging solution, combined with computer based automatic analysis approaches, is a powerful method of characterizing cell migration in different culturing conditions. Given these motivations, we have developed an automatic cell segmentation and tracking method targeted at low magnification phase-contrast microscopy images. The system is able to segment and track a large number of cells in confluent cultures.

**Exploitation of Results**

CMV is also a member of the Oulu BioImaging Network (OBI), which is a forum for promoting collaboration between the research groups and experts working in the bioimaging area at the University of Oulu and Oulu University Hospital. It is an associated partner to Euro-BioImaging, which aims at creating a coordinated and harmonized plan for the deployment of biomedical imaging infrastructure in Europe. In 2012, the first OBI Workshop was held at the Linnanmaa campus, where also the research activities of CMV were widely presented.
geometric calibration of Kinect with an external camera; this has received much interest from the other researchers worldwide. By the end of 2012, it has been downloaded over 5,000 times.

The results have been also utilized in our own projects. For example, we collaborate with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with physiological biosignals. Together with Prof. Seppänen and Dr. Seppo Lauka (Department of Educational Sciences and Teacher Education) and Prof. Matti Lehtihalmes (Faculty of Humanities) we are also participating in the FSR Second Wave project where we have developed a Mobile Multimodal Recording System (MORE) that will be used in classroom research in various schools.

Most of our funding for both basic and applied research comes from public sources such as the Academy of Finland and Tekes, but besides these sources, CMV also conducts research by contract which is funded by companies. In this way, our expertise is being utilized by industry for commercial purposes and even in consumer products, like mobile devices.

The CMV has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience, their roots are especially in the strands of “free academic research”. There are currently altogether five research based spin-outs founded directly on the machine vision area. The number of spin-outs could be extended up to sixteen when taking into account the influence of the CMV’s thirty-year old history and the spin-out companies from the spin-out research groups in the area of computer science and engineering in total.

**Future Goals**

In recent months we have put substantial effort into preparing our research and operation plan for the coming years. This was needed, for example, for the Finnish Centre of Excellence in Machine Vision Research proposal that we submitted in response to the Academy of Finland’s Call for Centre of Excellence Programme 2014–2019. From very tough competition, our proposal was selected for the 2nd stage to be evaluated in 2013. If accepted, our resources for carrying out well focused cutting-edge research, for example, on perceptual interfaces for face to face interaction, multimodal analysis of emotions, 3D computer vision, and energy-efficient architectures for embedded vision systems, would be significantly strengthened. We also have plans to further deepen our collaboration with international and domestic partners. For this purpose, we are participating in new European project proposals. Close interaction between basic and applied research has always been a major strength of our research unit. The scientific output of the CMV has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.

### Personnel

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors, doctors</td>
<td>19</td>
</tr>
<tr>
<td>doctoral students</td>
<td>18</td>
</tr>
<tr>
<td>others</td>
<td>6</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>43</strong></td>
</tr>
<tr>
<td><strong>person years</strong></td>
<td><strong>34</strong></td>
</tr>
</tbody>
</table>

### External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>752 000</td>
</tr>
<tr>
<td>Ministry of Education and Culture</td>
<td>121 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>411 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>175 000</td>
</tr>
<tr>
<td>international</td>
<td>201 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>1 660 000</strong></td>
</tr>
</tbody>
</table>

### Doctoral Theses


### Selected Publications


INFOTECH OULU Annual Report 2012 41


Background and Mission

The Center for Machine Vision Research (CMV) is a creative, open and internationally attractive research unit. It is renowned worldwide for its expertise in computer vision.

The center has a strong record, which now spans for over 32 years, of scientific merits on both basic and applied research on computer vision. It has achieved ground-breaking research results in many areas of its activity, including texture analysis, facial image analysis, geometric computer vision, and energy-efficient architectures for embedded systems. The mission of the center is to develop novel computer vision methods and technologies that create the basis for emerging innovative applications.

In February 2014, the CMV had three professors, one Associate Professor and one FiDiPro Professor, 16 senior or postdoctoral researchers, and 25 doctoral students or research assistants. The unit is highly international: 50% of our researchers (doctors, PhD students) are from abroad. CMV has an extensive international collaboration network in Europe, the USA, and China. The mobility of the researchers between leading research groups abroad, and vice versa, is intense. Within the Seventh Framework Programme FP7, the CMV currently participates in the project consortium of Trusted Biometrics under Spoofing Attacks (TABULA RASA). It also participates in two European COST actions.

The main areas of our research are computer vision methods, human-centered vision systems and vision systems engineering. The results of the research have been widely exploited in industry, and contract research forms a part of our activities.

Highlights and Events in 2013

In September, the Center for Machine Vision Research (CMV) was re-selected to Infotech Oulu for the period of 2014-2017, attaining the highest score among all applicant groups. In January 2014, the results of the Research Assessment Exercise (RAE 2013) of the University were released. An international panel, aided with a bibliometric analysis made by Leiden University, ranked CMV with the highest score 6 (outstanding), representing the international cutting edge in its field.

CMV played a significant role in the ninth Conferment Ceremony of the University of Oulu in May. Professor Matti Pietikäinen acted as the Conferrer of Degrees for the Faculty of Technology. Dr. Juho Kannala was chosen as Primus Young Doctor - the first young doctor whose degree was conferred in the ceremony - chosen according to his excellent merits. The University honored distinguished scientists or influential members of society with the award of an Honorary Doctorate. CMV core partner, Prof. Stan Z. Li (Chinese Academy of Sciences), received this award for his contributions to image pattern recognition and biometrics.

Two of the CMV senior researchers strengthened their role as independent team leaders. Dr. Guoying Zhao received a tenure track position as Associate Professor at the University of Oulu from January 2014 until the end of 2018. Dr. Abdenour Hadid received five-year Academy of Finland Research Fellow funding from September 2013 onwards. Both positions were highly competitive.

The IET Biometrics Premium Award 2013 was given to CMV. The Institution of Engineering and Technology (IET) presents one Premium Award per each for its journal to recognize the best research papers published during the previous year. CMV’s winning paper “Face spoofing detection from single images using texture and local shape analysis” is authored by Jukka Komulainen (né Määttä), Abdenour Hadid and Matti Pietikäinen. It was published in 2012 in the IET Biometrics journal, volume 1, issue 1. The winners were given a certificate at the 2013 IET Achievement Awards in London in November.

CMV has had a significant role in the TABULA RASA project, appraised as a success story by the European Commission and followed by a large media campaign. The consortium comprises 12 different organizations across seven countries that have worked together over a period of three years in improving the security of biometric systems. CMV has provided expertise in both face and gait recog-
nition using Local Binary Patterns in developing ways to detect the spoofing attacks. The same LBP methodology has also been utilized by other TABULA RASA partners. In addition, CMV has led the work package in evaluating the vulnerabilities in current biometric systems.

CMV contributed significantly to the 18th Scandinavian Conference on Image Analysis (SCIA 2013), organized in June in Espoo, Finland. The CMV Leader, Prof. Matti Pietikäinen co-chaired the SCIA conference with Prof. Erkki Oja. CMV Vice Leader, Prof. Janne Heikkilä acted as one of the area chairs. Prof. Pietikäinen and Dr. Guoying Zhao lectured in a tutorial “Image and video analysis with local binary patterns”. CMV researchers presented altogether as many as ten papers.

Scientific Progress

The current main areas of research are: 1) Computer vision methods, 2) Human-centered vision systems, and 3) Vision systems engineering. In addition to these main areas, our research activities include biomedical image analysis where we collaborate with Biocenter Oulu.

Computer vision methods

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. In the last few years, the research in computer vision methods has been broadened to cover also two other areas: computational photography, and object detection and recognition. The aim in all these areas is to create a methodological foundation for the development of new vision-based technologies and innovations.

Texture analysis

Texture is an important characteristic of many types of images and can play a key role in a wide variety of applications of computer vision and image analysis. The CMV has long traditions in texture analysis research, and ranks among the world leaders in this area. The Local Binary Pattern (LBP) texture operator has been highly successful in numerous applications around the world, and has inspired plenty of new research on related methods, including the blur-insensitive Local Phase Quantization (LPQ) method, also developed at CMV.

We proposed a simple and robust local descriptor, called the robust local binary pattern (RLBP). The basic LBP works successfully in many domains, such as texture classification, human detection and face recognition. However, an issue of LBP is that it is not so robust to the noise present in the image. We improved the robustness of LBP by changing the coding bit of LBP. Experimental results on the Brodatz and UIUC texture databases show that RLBP outperforms other widely used descriptors (e.g., SIFT, Gabor, MR8 and LBP) and other variants of LBP (i.e., completed LBP), especially when we add the noise in the images. In addition, experimental results on human face recognition also show a promising performance, comparable to the best known results on the Face Recognition Grand Challenge (FRGC) dataset.

Recently, local quantized patterns (LQP) was proposed for using vector quantization to code complicated patterns with a large number of neighbors and several quantization levels. It uses a lookup table technique to map patterns into the corresponding indices. Since LQP only considers the sign-based difference, it misses some discriminative information. We proposed completed local quantized patterns (CLQP) for texture classification. The magnitude and orientation-based differences are utilized to complement the sign-based difference for LQP. In addition, vector quantization is exploited to learn three respective codebooks for local sign, magnitude and orientation patterns. For reducing the unnecessary computational time of initialization, we used preselected dominant patterns as the initialization in vector quantization. Our experimental results show that CLQP outperforms well-established features, including LBP, LTP, CLBP and LQP on a range of challenging texture classification problems and an infant pain detection problem.

For improving the accuracy of LBP-based operators by including texture image intensity characteristics in the operator, we proposed the utilization of a shifted step function to minimize the quantization error of the step function to obtain more discriminative operators. Features obtained from the shifted step function are simply fused together to form the final histogram. This model is generalized, and can be integrated with other existing LBP variants to reduce quantization error of the step function for texture classification. The proposed method is integrated with multiple LBP-based feature descriptors, and evaluated on publicly available texture databases (Outex TC_00012 and KTH-TIPS2b) for texture classification. Experimental results demonstrate that it not only improves the performance of operators with which it is integrated, but it also compares favorably to the state of the art in texture classification.
Dynamic texture (DT) is an extension of texture to the temporal domain. How to segment DTs is a challenging problem. We addressed the problem of segmenting DT into disjoint regions. DTs might be different from their spatial mode (i.e., appearance) and/or temporal mode (i.e., motion field). To this end, we developed a framework based on the appearance and motion modes. For the appearance mode, we use a new local spatial texture descriptor to describe the spatial mode of DT; for the motion mode, we use the optical flow and the local temporal texture descriptor to represent the temporal variations of DT. In addition, for the optical flow, we use the Histogram of Oriented Optical Flow (HOOF) to organize them. To compute the distance between two HOOFs, we developed a simple, effective and efficient distance measure based on Weber Law. Furthermore, we also addressed the problem of threshold selection by proposing a method for determining thresholds for the segmentation method by offline supervised statistical learning. Experimental results show that our method provides very good segmentation results compared to the state-of-the-art methods in segmenting regions that differ in their dynamics.

Illustration of DTs: (a) DTs are different in their spatial mode (i.e., appearance) but show a similar temporal mode (i.e., motion), (b) DTs are different in their temporal mode but show a similar spatial mode, and (c) the similar temporal/spatial mode of DTs are cluttered.

Video texture synthesis is the process of providing a continuous and infinitely varying stream of frames, which plays an important role in computer vision and graphics. However, a challenging problem remains in generating high quality synthesis results. Considering the two key factors that affect the synthesis performance - frame representation and blending artifacts - we improved the synthesis performance from two perspectives: first, effective frame representation is designed to capture both the image appearance information in the spatial domain and the longitudinal information in the temporal domain. Second, artifacts that degrade the synthesis quality are significantly suppressed based on a diffeomorphic growth model. The proposed video texture synthesis approach has two major stages: the video stitching stage and the transition smoothing stage. In the first stage, a video texture synthesis model is proposed to generate an infinite video flow. To find similar frames for stitching video clips, we presented a new spatial-temporal descriptor to provide effective representation for different types of dynamic textures. In the second stage, a smoothing method is proposed to improve synthesis quality, especially from the point of view of temporal continuity. It aims to establish a diffeomorphic growth model to emulate local dynamics around stitched frames. The proposed approach is thoroughly tested on public databases and videos from Internet, and is evaluated in both qualitative and quantitative ways.

Overview diagram of the whole DT synthesis method, which consists of two main steps: video stitching using a Multiframe LBP-TOP signature (left), and transition smoothing using deformable image registration and growth model estimation (right).

Computational photography

In computational photography, the aim is to develop techniques for computational cameras that give more flexibility to image acquisition and enable more advanced features to be employed, going beyond the capabilities of traditional photography. These techniques often involve use of special optics and digital image processing algorithms that are designed to eliminate the degradations caused by the optical system and viewing conditions. In our recent work, new imaging solutions such as microlens array based sensing technology, and novel algorithms for facilitating image acquisition, coding and reconstruction have been the key areas of interest.

In our research related to light field imaging, we have investigated techniques for compression of light field images. In particular, our work has addressed an asymmetric scenario where the encoder should have low computational complexity, allowing its implementation on resource-limited devices. In the first phase, we have surveyed potential techniques, such as sampling approaches, image transforms, distributed source coding and compressed sensing techniques for the purpose.

In addition to the further development of microlens array based detection of parasites on a mobile platform, the
automatic parasite detection was studied in collaboration with the Institute for Molecular Medicine Finland and the Karolinska Institute. The on-chip imaging was successfully used to detect Schistosoma haematobium eggs. The future studies will be focused on improving the result image quality for data captured with the microlens array cameras and data captured with in-line holography.

Image registration is one of the most important and most frequently discussed image processing topics in the literature, and it is a crucial preprocessing step in all image analysis tasks in which the final information is obtained from a combination of various data sources, like image fusion, change detection, multichannel image restoration, superresolution, etc. In many cases, the images to be registered are inevitably blurred. The blur may originate from camera shake, scene motion, inaccurate focus, atmospheric turbulence, sensor imperfection, low sampling density and other factors. We developed an original registration method designed particularly for registering blurred images. Our method works for unknown blurs, assuming the PSF’s exhibit N-fold rotational symmetry. We proved experimentally its good performance which is not dependent on the amount of blur.

Two images blurred by two different symmetric PSF’s (left and middle). The result of multichannel blind deconvolution after registering the two images with our method (right).

**Object detection and recognition**

Just by glancing at an object, for example an apple or a building, a human is immediately aware of many of the object qualities. For instance, the apple may be red or green, and the building may be made of concrete or wood. These properties can be used to describe the objects and further qualify them. Currently, even the best systems for artificial vision have a very limited understanding of objects and scenes. For instance, state-of-the-art object detectors model objects as distributions of simple features (e.g., HOG or SIFT), which capture a blurred statistics of the two-dimensional shape of the objects. Color, material, texture, and most of the other object attributes are likely ignored in the process. Fine grained object classification and attributes have recently gained a lot of attention in computer vision, but the field is still in its infancy. For instance, currently there are only a few small databases.

Our research objective is to develop novel methods to reliably extract a diverse set of attributes from images, and to use them to improve the accuracy, informativeness, and interpretability of the object models. The goal is to combine advances in discrete-continuous optimization, machine learning, and computer vision, to significantly advance our understanding of visual attributes and produce new state-of-the-art methods for their extraction. We do this in three ways: by developing learning approaches, which utilize mid-level image segments to automatically find the combination of object parts that correspond to, possibly small, differences between two object classes (e.g. two bicycle models); by utilizing dependencies for learning complex attribute combinations using structured output models; by using crowd sourcing tools to discover a comprehensive vocabulary that is used by humans to describe objects when performing a particular task (e.g. browsing bicycle catalogue).

In 2012, CMV researchers participated in a workshop led by Prof. A. Vedaldi at Johns Hopkins University. As a part of the workshop, we began to collect a new extensive dataset that is intended to serve as a benchmark for detailed object attribute and part recognition. Part of this data was published in the Fine-Grained Visual Categorization workshop, organized in conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR) 2013. This data also became a part of the ImageNet FGVC challenge, which was in conjunction with the International Conference on Computer Vision (ICCV) 2013.

Efficiency is one of the key issues for real-time object detection. Although nonlinear classifiers are more powerful than the linear ones, few existing methods integrate them as the weak classifiers into the commonly used boosting framework. The reason mainly lies in that the conventional nonlinear classifiers usually have high computational costs. To address this problem, we proposed an efficient nonlinear weak classifier, named the Partition Vector weak Classifier (PVC). PVC is a weighted combination of a series of additive kernel functions of the (input) feature vector with respect to a set of pre-defined vectors, namely the Partition Vectors (PVs). The PVC’s learning includes three key steps: encoding, hyper-plane learning, and decoding. The obtained classifiers can be further accelerated using piecewise constant functions, such that it ensures a computational cost proportional to the dimension of the features during evaluation, as do the conventional linear classifiers.

PVC learning. (a) Encoding maps samples in the original space to the implicit space. (b) Learning the hyper-plane using the encoded samples in the implicit space. (c) Decoding transforms the learnt hyper-plane to the original space.

We demonstrated our algorithm in detection tasks for multiple classes of objects, including pedestrians, cars, bicycles, and cows, as illustrated below. Experimental results show that the boosted PVC significantly improves both the learning and evaluation efficiency of nonlinear SVMs to the level of boosted linear classifiers, without losing any of the high discriminative power.
Geometric computer vision

Images are 2D projections of the 3D world, which makes inferring 3D information an ill-posed problem from a single viewpoint, and a challenging problem from multiple views. Geometric computer vision provides the tools for establishing the relationship between the image and the 3D scene. While the fundamental theory of geometric computer vision has been developed already in the previous century, still for example, automatic construction of 3D scene models from multiple photographs is a relevant problem, and is subject to active research. Furthermore, new depth cameras, such as the Kinect sensor, have boosted rapid progress in scene modeling. Intelligent machines that require 3D information from the environment are a natural application area for geometric computer vision. Another application area that has gained much attention in the last few years is mixed reality, where real and virtual objects co-exist in the same environment. Wearable computers such as Google Glass have created a strong demand for such technology. Mixed reality has been also the key driver in our recent work on geometric computer vision.

During the last year, we have been developing a method for creating reconstructions from multiple photographs. Our previous method published in ICPR 2012, which basically takes a set of images as input, and outputs a point cloud in three-dimensional space, was extended with a couple of improvements. The improvements make the point clouds both denser and more accurate, without notable loss in computational efficiency. Hence, compared with the state of the art, our method produces reconstructions of similar or better quality, and is significantly faster. During the year, we also studied methods for generating triangular meshes from point clouds. The goal in this part is to turn a point cloud into a compact and watertight mesh of connected triangles so that it could be used, for example, to create virtual reality models. The pipeline from a set of images to a triangular mesh is illustrated in the figure below.

Establishing point-to-point correspondence between images is a fundamental problem in many applications such as augmented reality. In the past decade, many local image detection/description techniques have been developed to detect locations in images that are suitable for matching and to describe the visual properties of those points using the local image region around them. Robustness and computational efficiency are the two main criteria for choosing a particular local descriptor for an application. The need for real-time speed on video stream data has led to the emergence of many fast descriptors (Random Ferns, ORB etc) using simple pixel level comparisons which can be computed and matched efficiently. But they are not robust against camera pose variation. Robust features like SIFT/SURF have been successfully used under many challenging conditions. These robust features are computationally expensive. Our recent study aims at exploring different ways of accelerating the point-to-point correspondence of a matching process involving robust descriptors. Our initial experiments with SIFT vs. ORB have provided some hope of achieving fast point matching using SIFT descriptors while maintaining a superior accuracy over ORB descriptors. We are in the process of expanding the scope of the experiments to larger datasets involving widely varying scenarios.
Human-centered vision systems

In future ubiquitous environments, computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems, enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

Face recognition and biometrics

CMV continued playing a key role in the FP7 EU project TABULA RASA which has recently been selected as a success story by the European commission. TABULA RASA aims at researching, developing and evaluating countermeasures for spoofing attacks against biometric systems. In this context, we have proposed and evaluated advanced countermeasures for face and gait biometric modalities. We also co-organized, jointly with UNICA from the University of Cagliari (Italy), a spoofing challenge at the International Conference on Biometrics (ICB 2013) which was held in Madrid in June 2013. The aim of the challenge was to demonstrate the effects of spoofing and anti-spoofing in fingerprint biometrics and to raise awareness of the spoofing threats to biometric systems.

An example of a mask that can be used to attack (spoof) a biometric system.

We hence continued exploring promising directions for face spoofing detection within the context of the TABULA RASA project, in addition to analysis of facial texture and motion patterns that have shown to be effective in our previous studies. As we humans rely mainly on scene and context information when performing spoofing detection, we have been investigating approaches for exploiting contextual information in face anti-spoofing. In our initial studies, histogram of oriented gradients (HOG) descriptors were used for describing distinctive discontinuities around the detected face and determining whether a natural upper-body profile or the boundaries of the spoofing medium is detected in the scene. The proposed countermeasure improved the state of the art and showed promising generalization capabilities also in cross-database evaluation. Moreover, it is reasonable to assume that no single superior technique is able to detect all known, let alone unseen, spoofing attacks. Therefore, we have been studying how different countermeasures could be combined in order to construct a flexible network of attack-specific spoofing detectors in which new techniques can be easily integrated to patch the existing vulnerabilities in no time when new countermeasures appear. Together with the IDIAP Research Institute in Switzerland, we developed and published an open-source face anti-spoofing framework that includes several countermeasures and strategies for combining them. The same software framework was also successfully utilized in the 2nd Competition on Counter Measures to 2D Face Spoofing Attacks organized within the context of ICB 2013.

We also continued our research on recognizing human demographics (e.g. age and gender) from facial images with emphasis on local binary patterns (LBP). The most significant achievement in this domain is a method called the LBP kernel density estimate. Our extensive experiments showed very promising results especially in human age estimation, but also in texture classification and face recognition. The proposed method can be seen as an alternative to the widely used histogram representation, and it has potential in situations where the number of all possible local binary patterns producible by any given LBP operator exceeds the number of pixels in the image. The method provides an efficient way for preventing sparsity, which is a common problem with LBP histograms. The method also turned out to perform well with other LBP variants, for example, with CLBP, which is among the most powerful LBP variants.

Examples of automatically estimated age categories (ground truth in parantheses).

Recognition of facial expressions and emotions

The face is the key component in understanding the emotions, and this plays significant roles in many areas, from security and entertainment to psychology and education.

We proposed a method to detect facial action units in 3D face data by combining novel geometric properties and a new descriptor based on the Local Binary Pattern (LBP) methodology. The proposed method enables person and...
gender independent facial action unit detection. The decision level fusion is used by employing the Random Forests classifiers to combine geometric and LBP based features. Unlike the previous methods, which suffer from the diversity among different persons and normalize features utilizing neutral faces, our method extracts features on a single 3D face data. Besides, we show that an orientation based 3D LBP descriptor can be implemented efficiently in terms of size and time without degrading the performance. We tested our method on the Bosphorus database, and presented comparative results with the existing methods. Our approach outperformed existing methods, achieving a mean receiver operating characteristic area under curve (ROC AuC) of 97.7%.

Facial expression recognition (FER) has been predominantly utilized to analyze the emotional status of human beings. In practice, nearly frontal-view facial images may not be available. Therefore, a desirable property of FER would allow the user to have any head pose. Some methods on non-frontal-view facial images were recently proposed to recognize the facial expression by building a discriminative subspace in specific views. These approaches ignore (1) the discrimination of inter-class samples with the same view label and (2) the closeness of intra-class samples with all view labels. We proposed a new method to recognize arbitrary-view facial expressions by using discriminative neighborhood preserving embedding and multi-view concepts. It first captures the discriminative property of inter-class samples. In addition, it explores the closeness of intra-class samples with an arbitrary view in a low-dimensional subspace. Experimental results on BU3DFE and Multi-PIE databases showed that our approach achieves promising results for recognizing facial expressions with arbitrary views.

It is commonly agreed that emotions are a multimodal procedure. Combining complementary information from different modalities may increase the accuracy of emotion recognition. In the AFFECT project, funded by TEKES, we have been investigating the fusion of different modalities e.g., spontaneous facial expressions as an external channel and electroencephalogram (EEG) as an internal channel, supplementing facial expressions for more reliable emotion detection in long continuous videos.

Analysis of visual speech

Human speech perception is a bi-modal process which makes use of information not only from what we hear (acoustic) but from what we see (visual). In machine vision, visual speech recognition (VSR) is the task of recognizing the utterances through analyzing the visual recordings of a speaker’s talking mouth without any acoustic input. Although visual information cannot in itself provide normal speech intelligibility, it may be sufficient within a particular context when the utterances to be recognized are limited. In such a case, VSR can be used to enhance natural human-computer interactions through speech, especially when audio is not accessible or is severely corrupted.

Our research is focused on the extraction of a set of compact and informative visual features for VSR. To do that, the generative latent variable model is adopted to model the inter-speaker variations of visual appearances and those caused by uttering. Moreover, we propose to use a path graph to capture the temporal relationships of video frames. The low-dimensional continuous curve embedded within the graph is used as prior knowledge when constructing prior distributions of latent variables. Our method has been compared with the state-of-the-art visual features and has achieved superior results.

Illustration of multi-view discriminative neighborhood preserving embedding for arbitrary-view FER.
Visual speech can also be used for determining the identity of a person. A novel local spatiotemporal directional descriptor was proposed for speaker identification by analyzing mouth movements. For this new descriptor, the directional local binary pattern features in three orthogonal planes are coded. In addition, besides sign features, magnitude information encoded as weight for the bins with the same sign value is developed to improve the discriminative ability. Moreover, decorrelation is exploited to remove the redundancy of features. Experimental results on the challenging XM2VTS database show the effectiveness of the proposed representation for this problem.

Human tracking and action analysis

Even though much work has been done for action recognition, minor efforts have been dedicated to understanding emotion from analyzing action, e.g. people’s walking. We have collected an affective gait database and designed descriptors to be robust against rotation and scale variations that occur during recording gait data in the real world while individuals are truly affected emotionally.

In order to improve the user experience with a large touchscreen, we introduced gesture interaction based on a Kinect sensor in a wall-sized touchscreen. According to the distance between the user and the display, we created two interaction modes: ‘Near-Mode’ for touch interaction; ‘Far-Mode’ for gesture interaction. With this solution, the interaction is more user-friendly and young users or users in wheelchairs are also able to interact with the large touchscreen applications.

Affective human-robot interaction

Development of our experimental HRI platform, Minotaurus, was continued with the support of the European Regional Development Fund (2010-2014) in collaboration with the Intelligent Systems Group. Minotaurus consists of a Segway Robotic Mobility Platform (RMP 200) and a set of laptops, Kinect sensors, video cameras, laser scanners, microphones, magnetic field sensors, a robot arm, an avatar display and a ubiquitous multi-camera environment.

During the last year of the project, the work continued to integrate various components to function together. We also developed methods of controlling the robotic arm using observations from Kinect sensor. We have successfully demonstrated that Minotaurus can recognize various objects from a table surface, has the capability to plan how to pick up the detected object, and can execute the plan by controlling the robotic arm.
Minotaurus is also capable of detecting people from a distance and can understand some of their gestures. From a closer distance, it can detect and recognize familiar faces, analyze faces to detect facial expressions and the gender of the person. It can also understand both Finnish and English voice commands, and reply using spoken sentences with the same language. While the robot speaks, the mouth movements of the avatar are synthesized to match the generated speech. Minotaurus can also understand the environment and navigate to its target while avoiding obstacles by using a combined environment model generated from all the sensors.

Minotaurus and its capabilities have been successfully demonstrated at various private and public events like, for example, during the University Science Day and a robot-themed event at the science center Tietomaa. This way, the demonstrations have been arranged not only in our robotic lab, but in real environments, and the reactions of the audiences have been entirely positive and enthusiastic.

An overview of the perceptual system of Minotaurus and an illustration of the developed interaction modalities.

Vision systems engineering

Vision systems engineering research aims to identify attractive computing approaches, architectures, and algorithms for industrial machine vision systems. In this research, solutions ranging from low-level image processing even to equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial solutions. We have also applied our expertise to applications intended for embedded mobile platforms.

The framework for a lumber tracing system was developed using 1D projection signals together with local descriptors. The product identification theme was continued, and the next logical step in the wood refinement chain was to find the link between boards and log ends. This is a very challenging problem, where a small image patch within the log end has to be matched to the board end image. The log end can randomly rotate 360 degrees and the image patch can be located almost anywhere within the log end. As with the board side images, properly selected local descriptors have shown great potential for correct matching. An example log end – board end matching result is shown below.

In the research area of interactive mobile applications, we have studied multimodal gesture controlled user interaction. The methods developed utilize the multiple computing resources present on current mobile platforms such as GPUs. The gestures are recognized from the front camera and the touch screen. With the user interface, the user can move the mouse cursor, click on objects and scroll documents. The functions provided to the user depend on the distance between the hand and the device. For this purpose, we have developed a new finger detection and tracking system, based on color and motion features.

Geometrically consistent matches between the rotated log end image and the board end image.

Our work on motion-based object detection and tracking has also continued. The aim in this work is to integrate feature based sparse motion segmentation with a sampling based motion detection and tracking framework, which would lead to efficient solutions applicable in online dynamic scene analysis tasks. The method is designed for mobile platforms and can be utilized, for example, in gesture controlled user interfaces.

In the field of energy-efficient embedded computer vision, we have implemented several variants of the LBP operator in multiple mobile and custom processors. The embedded platforms used range from multicore-ARM and mobile GPUs to TTA processors and a hybrid SIMD/MIMD image co-processor. We have compared the different implementations in terms of computational performance and energy efficiency, while analyzing the different optimizations that can be made on each platform and its different available computing resources. In addition, we have released a software package providing a valuable tool for other researchers and developers.

Two computationally intensive multimedia applications - face detection and depth estimation - were implemented and optimized for parallel processing using the Portable computing language (PoCL) implementation of Open Computing Language (OpenCL). So far, the benchmarks have been implemented on desktop CPU and GPU. An ini-
tial design of an energy efficient multicore transport triggered architecture (TTA) processor that could achieve the same performance with significantly lower energy consumption has also been implemented, but not yet benchmarked.

The Energy Efficient Architectures and Signal Processing team of CMV has been working on design automation and energy efficient computing for signal processing applications. A remarkable new opening was the initiation of a joint US-Finnish research project CREAM, together with the Centre for Wireless Communications. During the first project year, the research focus has been on dataflow modeling and energy-efficient implementation of a digital pre-distortion filter for wireless mobile transmitters. One of our doctoral students, Amanullah Ghazi, also conducted a 2-month research visit to the University of Maryland on the basis of this project with Infotech financial support. Another 2-month research visit was made by Dr. Jani Boutellier to EPFL, Switzerland focusing on the topic of dataflow programming.

In the context of video processing, a programmable, energy-efficient processor for HEVC/H.265 adaptive loop filtering was developed. This work is also to be extended to further parts of the latest H.265 video compression standard. This research topic has been carried out, heavily supported by the Academy of Finland. In general, the project creates tools for generating efficient embedded software/hardware solutions. Platform independent high-level specifications are used to describe parallelism at data, instruction, task and memory levels. The target is many-core systems that are becoming the key approach in improving computing throughput.

Biomedical image analysis

In recent years, increasing resolving power and automation of microscopic image acquisition systems have resulted in an exponential increase in microscopic data set sizes. Manual analysis of these data sets is extremely labor intensive and hampers the objectivity and reproducibility of results. There is, therefore, a growing need for automatic image processing and analysis methods. Biomedical image analysis is an emerging application area in which we have collaborated with Biocenter Oulu for few years.

We have recently started a new project called “Algorithm-based combination and analysis of multidimensional video and open data” (ABCdata), funded by Tekes. In this project, the objective is to analyze 3D microscopic image sequences, and develop tools for cell segmentation and tracking, as well as for detection of cellular events such as mitosis and apoptosis in conditions that mimic human tissues, which makes this research unique from the scientific point of view.

One of the topics we have been investigating in the project during the last year is analysis of cancer progress. The ability of a cell to sense its environment and adapt to it and its morphological appearance is a crucial element in tumor progression. Therefore, analysis of these morphology changes and cell dynamic behavior in long term living cell imaging is a critical investigation in cell biology and drug development research. To monitor and quantify the cell dynamics in cancer biology in live cell microscopy, which is comprised of long image sequences, automated image analysis solutions are needed. Therefore we have been developing computer vision/ machine learning methods to address these needs. In this project, we employ phase contrast and fluorescent images taken from 3D models in which tumors reside and interact dynamically with the surrounding matrix and fibroblast cells. Recently, we proposed an automated method, based on a learning framework, for detecting tumor cells. The proposed method, which can be employed in different applications in biomedical image analysis, is able to distinguish different cell types in cell co-cultures and does not suffer from parameter tuning.

Sample images from our database. (Upper Left) Phase contrast image of a 3D culture containing tumor (roundish) and fibroblast cells (elongated). (Upper Right) Fluorescent image of the same culture. Green Fluorescent Protein (GFP) is used to label fibroblast cells. (Lower Left) Phase contrast and fluorescent images are superimposed and interpretation of the phase contrast image is easier. (Lower Right) Our learning based probabilistic tumor detection result. The colormap indicates confidence level.

Accurate cell segmentation is a prerequisite of any detailed analysis of microscopic images. Good segmentation results can greatly simplify many analysis tasks. During the previous year we have been working with GFP labeled squamous carcinoma cells (HSC-3) embedded in the 3D collagen matrix, and image stacks have been captured with a spinning disk confocal microscope. One of the major challenges with our data is separating cells that touch. Most approaches make some assumptions about the shape of the cells (usually cells are assumed to be round and their size to be within a narrow range). These approaches do not work well when cells have very flexible shapes and can have intensity variation within their body.

We have attempted to separate touching cells using a cascade of segmentation methods. Our method works better than basic segmentation methods alone, but it is still far from solving the difficult case of a dense cell sample with non-uniform intensity within the cell and very flexible cell shapes.
Phase-contrast illumination is simple and is the most commonly used microscopic method to observe non-stained living cells. Together with automatic cell segmentation and motion analysis tools, even single cell motility in large cell populations can be analyzed. To develop better automatic tools for analysis of low magnification phase-contrast images in time-lapse cell migration sequences, we have developed a segmentation method that relies on the intrinsic properties of maximally stable extremal regions (MSERs). In order to analyze cell migration characteristics in time-lapse movies, MSER-based automatic cell detection was combined with our own Kalman filter based multi-object tracker that efficiently tracked individual cells even in confluent cell populations. The research was conducted in cooperation with Biocenter Oulu and the University of Jyväskylä. The results have been reported in a joint article recently published in the Journal of Microscopy.

**Exploitation of Results**

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern methodology and its variants are used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis. The researchers in CMV have actively published the source codes of their algorithms for the research community, and this has increased the exploitation of the results.

The results have also been utilized in our own projects. For example, we have collaborated with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with physiological biosignals. Together with Prof. Seppänen and Dr. Seppo Laukka (Department of Educational Sciences and Teacher Education) and Prof. Matti Lehtihalmes (Faculty of Humanities) we have participated in the FSR Second Wave project where we have developed a Mobile Multimodal Recording System (MORE) that is now actively used in classroom research in various schools.

Most of our funding for both basic and applied research comes from public sources such as the Academy of Finland and Tekes, but besides these sources, CMV also conducts research by contract which is funded by companies. In this way, our expertise is being utilized by industry for commercial purposes, and even in consumer products, like mobile devices.

The CMV has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience, their roots are especially in the strands of “free academic research”. There are currently altogether five research based spin-outs founded directly on the machine vision area. The number of spin-outs could be extended up to sixteen when taking into account the influence of the CMV’s thirty-year old history and the spin-out companies from the spin-out research groups in the area of computer science and engineering in total.

**Future Goals**

The very positive results obtained from the RAE 2013 and Infotech Oulu evaluations show that we are on the right track. We plan to carry out well focused cutting-edge research, for example, on novel image and video descriptors, perceptual interfaces for face to face interaction, multimodal analysis of emotions, 3D computer vision, and energy-efficient architectures for embedded vision systems. We also have plans to further deepen our collaboration with international and domestic partners. For this purpose, we are participating in new European project proposals. Close interaction between basic and applied research has always been a major strength of our research unit. The scientific output of the CMV has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.

**Personnel**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>professors, doctors</td>
<td>20</td>
</tr>
<tr>
<td>doctoral students</td>
<td>20</td>
</tr>
<tr>
<td>others</td>
<td>5</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>45</td>
</tr>
<tr>
<td><strong>person years</strong></td>
<td>39</td>
</tr>
</tbody>
</table>

**External Funding**

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>834 000</td>
</tr>
<tr>
<td>Ministry of Education and Culture</td>
<td>198 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>732 000</td>
</tr>
<tr>
<td>domestic private</td>
<td>80 000</td>
</tr>
<tr>
<td>international</td>
<td>102 000</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>1 946 000</td>
</tr>
</tbody>
</table>
Doctoral Theses


Selected Publications


Li X, Chen J, Zhao G & Pietikäinen M (2014) Remote heart rate measurement from face videos under realistic situations. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, in press.


Background and Mission

The Center for Machine Vision Research (CMV) is a creative, open and internationally attractive research unit. It is renowned world-wide for its expertise in computer vision.

The Center has a strong record, which now spans for over 33 years, of scientific merits on both basic and applied research on computer vision. It has achieved ground-breaking research results in many areas of its activities, including texture analysis, facial image analysis, geometric computer vision, and energy-efficient architectures for embedded systems. The mission of the center is to develop novel computer vision methods and technologies that create basis for emerging innovative applications.

In February 2015, the staff of CMV consists of three Professors, one Associate Professor, one FiDiPro Professor, and 16 senior or postdoctoral researchers (including three Academy Research Fellows), as well as 28 doctoral students or research assistants. The unit is highly international: over 50% of our researchers (doctors, PhD students) come from abroad. CMV has an extensive international collaboration network in Europe, China and the USA. Both outgoing and incoming mobility of researchers is intense to/from leading research groups abroad. Within the Seventh Framework Programme FP7, the CMV participated in the project consortium of Trusted Biometrics under Spoofing Attacks (TABULA RASA). It also participates in two European COST actions.

Scientific Progress

The current main areas of research are: 1) Computer vision methods, 2) Human-centered vision systems, 3) Vision systems engineering, and 4) Biomedical image analysis.

Highlights and Events in 2014

In January, the results of the Research Assessment Exercise (RAE 2014) of the University of Oulu were announced. An international panel, aided with a bibliometric analysis made by the Leiden University, ranked CMV with the highest score 6 (outstanding), representing international cutting edge in its field.

In September, CMV was awarded with the 2014 Jan Koenderink Prize for fundamental contributions in computer vision. The prize is handed out biennially at one of the most prestigious conferences in the field, the European Conference on Computer Vision, for a paper published exactly ten years ago at this venue which has stood the test of time.

CMV awarded with the Jan Koenderink Prize at ECCV 2014 in Zurich.

The awarded publication “Face Recognition with Local Binary Patterns” was authored by our alumni, Dr. Timo Ahonen (now with Nokia Technologies), Adj. Prof. Abdenour Hadid and Prof. Matti Pietikäinen. Automatic face recognition with Local Binary Patterns, a methodology developed at the University of Oulu, is currently regarded as one of the major milestones in face recognition research. The conference publication and its extension to a journal article have already been cited over 3,600 times according to Google Scholar.

At the same venue, the European Conference on Computer Vision (ECCV 2014), CMV contributed to three workshops. Assoc. Prof. Guoying Zhao and Prof. Matti Pietikäinen co-organized a workshop on Spontaneous Facial Behavior Analysis together with Prof. Stefanos Zafeiriou and Prof. Maja Pantic from Imperial College London. Adj. Prof. Abdenour Hadid co-hosted two workshops: the second International Workshop on Computer Vision with Local Binary Pattern Variants (LBP2014) jointly with Prof. Jean-Luc Dugelay, Eurecom and Prof. Stan Z. Li, Chinese Academy of Sciences; and another on Soft Biometrics jointly with Ass. Prof. Paulo Lobato Correia, University of Lisbon and Prof. Thomas Moeslund, Aalborg University.
As well, Dr. Jie Chen, Assoc. Prof. Guoying Zhao and Prof. Matti Pietikäinen co-organized a workshop on RoLoD: Robust Local Descriptors for Computer Vision, in conjunction with the ACCV 2014 conference. Now they are co-editing a special issue for the Neurocomputing journal

CMV succeeded very well in the latest call for funding of the Academy of Finland. Adj. Prof., Academy Research Fellow Abdenour Hadid obtained a two-year project funding from the first call of the ICT 2023 research program. His project aims at novel solutions for robust audio-visual biometrics. In addition, a four-year project on continuous emotional state analysis, led by Assoc. Prof., Academy Research Fellow Guoying Zhao, got accepted within the Academy’s general call for projects. Further, Dr. Juho Kannala obtained an Academy Research Fellow position and Dr. Jani Boutellier funding for postdoctoral research.

As many as four senior researchers of CMV were appointed as Adjunct Professors (docents) at the University of Oulu during the year. Dr. Jani Boutellier; Dr. Esa Rahtu; Academy Research Fellow, Dr. Juho Kannala; and Acting Professor, Dr. Jari Hannukela have all now proofed to possess comprehensive knowledge of their field of science along with a capacity for independent research and good teaching skills.

In May, CMV was recognized as one of “100 Acts from Oulu”. This campaign highlights actions that take the region of Oulu forward. The selection panel stated that research in machine vision is internationally renowned, and the results have attracted experts from all over the world to Oulu. This way, CMV’s research has made Oulu a more international city.

Computer Vision Methods

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. In the last few years, the research in computer vision methods has been broadened to cover also two other areas: computational photography, and object detection and recognition. The aim in all these areas is to create a methodological foundation for the development of new vision-based technologies and innovations.

Texture Analysis

Texture is an important characteristic of many types of images and can play a key role in a wide variety of applications of computer vision and image analysis. The CMV has long traditions in texture analysis research, and ranks among the world leaders in this area. The Local Binary Pattern (LBP) texture operator has been highly successful in numerous applications around the world, and has inspired plenty of new research on related methods, including the blur-insensitive Local Phase Quantization (LPQ) method, Weber Law Descriptor (WLD), and Binarized Statistical Image Features (BSIF), also developed by researchers of CMV. The first paper on LBP was published two decades ago at ICPR 1994 conference. A survey on recent progress of LBP was written for Prof. Erkki Oja’s honorary book published by Elsevier in 2015.

Effective characterization of texture images requires exploiting multiple visual cues from the image appearance. The local binary pattern (LBP) and its variants achieve great success in texture description. However, because the LBP(-like) feature is an index of discrete patterns rather than a numerical feature, it is difficult to combine the LBP(-like) feature with other discriminative ones by a compact descriptor. To overcome the problem derived from the nonnumerical constraint of the LBP, we proposed a numerical variant accordingly, named the LBP difference (LBPD). The LBPD characterizes the extent to which one LBP varies from the average local structure of an image region of interest. It is simple, rotation invariant, and computationally efficient. To achieve enhanced performance, we combine the LBPD with other discriminative cues by a covariance matrix. The proposed descriptor, termed the covariance and LBPD descriptor (COV-LBPD), is able to capture the intrinsic correlation between the LBPD and other features in a compact manner. Experimental results show that the COV-LBPD achieves promising results on publicly available data sets.

Deep learning is currently providing a major impact on computer vision research. The CMV is also investigating its usefulness in problems of its interest. For dynamic texture and scene classification, we have developed an approach, in which we train a deep structure by transferring some prior knowledge from image domain to video domain. Excellent results are obtained for three different benchmark datasets.

Computational Photography

In computational photography, the aim is to develop techniques for computational cameras that give more flexibility to image acquisition and enable more advanced features to be employed, going beyond the capabilities of traditional photography. These techniques often involve use of special optics and digital image processing algorithms that are designed to eliminate the degradations caused by the optical system and viewing conditions.

The meteorological visibility estimation is an important task for example in traffic control and aviation safety, but variable lighting conditions make this a challenging task to automate. Night time scenes with urban surroundings have a dynamic range up to seven decades while digital cameras typically have a dynamic range of four or five decades which clearly is not enough.

We have developed a computational imaging method enabling scene visibility classification during day and night. The high dynamic range imaging based on image stacks enables producing images that are very similar despite the time of the day. Retinex filtering.
that is high pass filtering in logarithmic space, further reduces the effects of lighting changes. Feature vectors extracted from the Retinex filtered HDR-images enabled automated visibility classification with 85.5% accuracy.

A scene captured mid-August in Oulu at midnight (left) and at midday (right). The tone mapped versions of HDR data on the top row show very little differences in the image contents which can be verified also on the Retinex filter response on the bottom row. The feature vectors used in classification (not shown) are projections of the of filtered images on y-axis.

In 2014 an ITEE Collaboration project was carried out with the Optoelectronics and Measurement techniques Laboratory aiming for inkjet 3D printed lenses for direct light field imaging. The initial results show that the quality of new inkjet printed lenses is much better than the ones previously produced with hot embossing. While there is still some development to be done, there are indications that the image quality might be pushed to match the reference glass lenses.

In-line holography enables particle measurements in large imaging volumes with an extended depth of field compared with conventional imaging systems. The accurate measurements of the structural details of the particles are practically possible only if the measured details are brought in focus. In-line holograms produce a stack of 2D images where the objects in focus have sharp features but out of focus objects and dual image present inherently in the in-line holography data, introduce extra noise making the focusing task challenging. We developed a new depth estimation method where the stack of reconstructed intensity images is analyzed. First rough object locations are estimated and the object depths are extracted with a wavelet based focus measure. Clusters of depth estimates are used with plane fitting to approximate the object orientation in the 3D volume to obtain the final all-in-focus images.

Image registration is one of the most important and most frequently discussed topic in the image processing literature, and it is a crucial preliminary step in all the algorithms in which the final result is obtained from a fusion of several images, e.g. multichannel image deblur, super-resolution, depth-of-field extension, etc. In many cases, the images to be registered are inevitably blurred. We developed an original registration method designed particularly for registering blurred images. Our method is specifically designed to work with images degraded by blurs modeled by point-spread-functions having dihedral symmetry, i.e. having both rotational and axial symmetry. Our registration algorithm is particularly well-suited in applications where common registration methods fail, due to the amount of blur. We proved experimentally its good performance which is independent from the amount of blur contained in the images.

Two images acquired with different focus settings and blurred by dihedrally symmetric PSF's (top-left and top-right). The PSF of the camera (bottom-left). The result of depth-of-field extension on after registering the two images with our method (bottom-right).

Object Detection and Recognition

Nowadays object detection and recognition is an important research area in computer vision as it has potential to facilitate search from very large unannotated image databases. However, even the best current systems for artificial vision have a very limited understanding of objects and scenes. For instance, state-of-the-art object detectors model objects as distributions of simple features (e.g., HOG or SIFT), which capture a blurred statistics of the two-dimensional shape of the objects. Color, material, texture, and most of the other object attributes are likely ignored in the process. Fine grained object classification and attributes have recently gained a lot of attention in computer vision, but the field is still in its infancy. For instance, currently there are not many databases, which would facilitate learning fine-grained object attributes. In order to alleviate this problem, we have collaborated with other researchers and contributed to the OID:Aircraft dataset which contains airplane images with detailed annotations for fine grained visual categorization of aircraft.

A successful recent paradigm in object detection is to replace traditional sliding window based approaches with so called object detection proposals, which can be considered as candidate regions on which category-specific object classifiers are evaluated. We have also studied object detection proposals and developed a method that combines local and global segmentation techniques for proposal generation.
Finally, related to object tracking, we have studied an approach, where motion based segmentation is integrated with occlusion based depth order analysis. The work builds on our earlier work on feature based video segmentation, where segmentation information is propagated from frame to frame using motion compensation. The ultimate goal of the study is to provide a computationally efficient method for detecting foreground objects, which is applicable to online moving camera applications such as camera-based user interfaces.

**Automatic Recognition of Movie Characters**

Television broadcasting has seen a paradigm shift in the last decade, as the Internet has become an increasingly important distribution channel. Delivery platforms, such as the BBC’s iPlayer, and distributors and portals like Netflix, Amazon Instant Video, Hulu, and Youtube, have millions of users every day. These new platforms include search tools for the video material based on the video title and metadata -- however, for most part it is not possible to search directly on the video content, e.g. find clips where a certain actor appears. Enabling such search services requires annotating the video content, and this is the goal towards which we work on.

Our objective is to automatically cluster and classify face tracks throughout a broadcast according to identity, i.e. to associate all the face tracks belonging to the same person. We investigate this problem in two different setups, where the first one does not assume any additional information to be available, and the second one utilises subtitles and transcripts that are often easy to obtain for a particular film. In literature, these two cases are often referred as unsupervised and weakly supervised approaches.

In the unsupervised case, the goal is to form links between all face tracks belonging to the same person. If successful, then annotating the video content for all actors requires simply attaching a label (either manually or automatically) for each cluster. To this end, the novelty in our research is to take account of the editing structure of the video when considering candidate face tracks to cluster. These cues have not been used in previous works. In particular we show that the shot-thread structure can be very useful cue to extract situations where two face tracks should or should not be assigned to the same person. Figure below shows a few examples where such information can be obtained from the threading pattern.

If the textual cues from subtitles and transcripts are available, we can utilise them to automatically mine the actor identities. This gives several benefits over the unsupervised case. For instance, we can directly name the face tracks instead of just clustering them. Furthermore, the textual cues allow us to make links between face tracks that are substantially different visually and hence difficult for the unsupervised methods. The main problem with subtitles and transcripts is that they do not contain information of the background characters and therefore it is very challenging to name them correctly. One of the main novelties in our research is a new efficient approach for handling the background character class. In this way we have been able to demonstrate a significant improvements over the previous state-of-the-art methods.

![Overview of the video-editing structure of a typical TV series episode. We see face tracks in a shot (top row), in a threading pattern (middle row) and in a scene (bottom row). Face tracks with the same color denote the same person. Must-not links between tracks are denoted by red edges.](image)

Examples of background characters. Each row shows three frames from one face track.

**3D Computer Vision**

In recent years, 3D computer vision has been subject to active research both in academia and industry. One of the key drivers has been the huge progress in range camera technology, and introduction of low cost sensors such as Microsoft Kinect. Despite of the many benefits provided by range cameras they are still mainly constrained to work in indoor conditions with limited operating range. Therefore, conventional 3D reconstruction from multiple photographs is still a highly relevant research topic. While the fundamental theory of geometric computer vision has been already developed several decades ago, using it in real application problems is an active area of research where novel...
scientific contributions and engineering work are needed.

3D computer vision has been one of the core research areas in CMV since early 1990’s. This research has resulted in many novel methods and software tools that have been widely used in the research community and companies. During the last few years our focus has been in 3D computer vision techniques that enable more advanced features in augmented reality applications, where real and virtual objects co-exist in the same environment. Wearable computers such as Google Glass have created a strong demand for such technology. One of the fundamental problems investigated in our work is accurate localization of the user with respect to the environment. Other important research problems include estimation of the 3D scene structure and building a dense 3D model from multiple images.

Simultaneous localization and mapping is the process of recovering the structure of a scene and the position of a moving camera simultaneously in real-time. It is a key component that enables many advanced applications of computer vision. For example, it is the underlying technology necessary for augmented reality applications. In itself, a SLAM system integrates many aspects of computer vision (e.g. feature detection, feature matching, triangulation, bundle adjustment) as well as many engineering challenges (e.g. real-time constraints, multi-threading, visualization). We have developed a SLAM system that uses the available visual information more effectively by incorporating both triangulated and non-triangulated features into the pose estimation and bundle adjustment stages. The system itself is portable to different architectures and can be easily extended by design. Moreover, the developed system has been released as open source with a very permissive license for the community to build upon. We expect that this contribution will further community interaction and advance the speed of research by enabling researchers to build upon our work.

We have also developed a strategy to accelerate the process of aligning 3D lines of a model with the 2D lines in the image of the model using purely geometric properties. Exploring all the possible camera position and orientation in order to obtain highest number of matches between 3D and 2D lines is prohibitively expensive. By using the geometric constraints arising solely from rotation parameters we are able to preempt many unsuitable camera poses. This significantly accelerated the task of establishing correspondence between 3D model lines and 2D image lines. Besides this, we also eliminated a degenerate case from the state-of-art line based pose estimation algorithm by reformulating the rotation parameters.

During 2014, we have also been developing methods for creating and evaluating triangle meshes. In the triangle mesh creation, we have focused on lightweight and simplified meshes which are generated from dense point clouds. For minimizing the complexity of the mesh, the method exploits the partial planarity of certain, usually architectural scenes. Because of the lack of available benchmarks, the evaluation of the quality and completeness of produced results has only been visual. Therefore, we have also been developing a benchmark for such evaluations which will be published later together with the triangle mesh generation method.

**Human-Centered Vision Systems**

In future ubiquitous environments, computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems, enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

**Face Recognition and Biometrics**

The FP7 EU project TABULA RASA (2010-2014) has ended in April 2014. CMV has played a key role in this project which has been selected as a success story by the European commission. CMV has provided expertise in both face and gait recognition using Local Binary Patterns in developing ways to detect the spoofing attacks. The same LBP methodology has also been utilized by other TABULA RASA partners. In addition, CMV has successfully led the work package in evaluating the vulnerabilities in current biometric systems.

Continuing the anti-spoofing theme, we started a joint project with the Speech and Image Processing Unit (SIPU), School of Computing University of Eastern Finland, Joensuu, Finland. The aim is to study, develop and evaluate novel audio-visual biometric authentication solutions for enhancing information security, especially under spoofing attacks. This project combines the strengths and expertise of two research groups in
Finland (CMV and SIPU). The consortium has a strong background on research areas related to the project. The focus is on face and voice biometrics and targets in-depth research toward innovative bi-modal solutions.

Examples of voice and face attacks on biometrics systems.

Besides voice and face anti-spoofing, we have also studied iris anti-spoofing aiming at detecting contact lenses. Although iris is considered to be the most reliable and accurate biometric trait for person identification, iris based biometric systems are prone to sensor level attacks (spoofing and obfuscation) like systems using any other biometric modalities. CMV has introduced a novel approach towards generalized textured contact lens detection by extracting binarized statistical image features (BSIF) from Cartesian iris images. In iris image classification, the ring-shaped iris region is traditionally mapped into a rectangular image also when detecting counterfeit irises. While the polar domain representation is convenient for finding distinctive features across different individuals and matching purposes, the geometric transformation causes severe distortion on the regular lens texture patterns. Our findings support the intuition that the textural differences between genuine iris texture and fake ones are best described by preserving the regular structure of different printing signatures without transforming the iris images into polar coordinate system. The proposed BSIF based iris texture representation showed very promising generalization capabilities across unseen textured contact lens types and different iris sensors on the latest benchmark datasets.

Cropped iris images highlighting the variation in texture patterns between one genuine iris and three textured lens manufacturers, Cooper Vision, Johnson & Johnson and Ciba Vision, respectively.

We continued our research on recognizing human demographics (e.g. age and gender) from facial images with emphasis on local binary patterns (LBP), the most significant achievement being a unified framework for learning LBP-like local image descriptors. The framework can be used in both supervised and unsupervised modes, and it generalizes several previous local image descriptors that are based on binarization. In the paper ‘Learning local image descriptors using binary decision trees’ we evaluated our proposed framework using varying levels of supervision in the descriptor learning phase. Our extensive experiments showed very promising results especially in human age group classification, but also in texture classification. The main contribution is to provide ways to learn the binary tests instead of fixing them by hand, like in the standard LBP method.

CMV co-organized the First ECCV 2014 International workshop on Soft Biometrics which was held in Zurich on September 7th, 2014 in conjunction with the European Conference on Computer Vision (ECCV 2014). The event was co-organized by the COST ACTION IC 1106 dealing with ”Integrating Biometrics and Forensics for the Digital Age” and particularly by the Working Group WG3 focusing on ”Forensic Behavioural and Soft Biometrics”. The event was open to researchers from both COST and non-COST countries. This workshop provided a clear summary of the state-of-the-art and discussed the most recent developments in soft biometric research with applications to forensics, surveillance and identification. There were about 30-40 participants attending the workshop and listening to 12 oral presentations. Besides these 12 oral presentations, the workshop also included three excellent keynote speeches on “Bag of Soft Biometrics for Person Identification: new trends and challenges” by Prof. Jean-Luc Dugelay, Eurecom, France, “What is the potential of soft biometrics for forensic applications?” by Prof. Massimo Tistarelli, University of Sassari, Italy, and “Soft biometrics for surveillance” by Prof. Mark Nixon, University of Southampton, UK.

Recognition of Facial Expressions and Emotions

The face is the key component in understanding the emotions, and this plays significant roles in many areas, from security and entertainment to psychology and education.
The mis-alignment and large variations in the temporal sale of facial expressions are two crucial problems for facial expression recognition. For handing these challenging problems, we proposed to use variation of canonical correlation that has the robustness to variations of alignment. However, the original canonical correlation ignores the facial expression variations and temporal information of facial expressions. We revisited the canonical correlation and then proposed an improved canonical correlation by using the followed tricks: (1) it uses the local binary pattern to describe the appearance features for enhancing the spatial variations of facial expression; (2) it develops the temporal orthogonal locality preserved projection for building a canonical subspace of a video clip, where it mostly captures the motion changes of facial expressions; and (3) it uses Fisher criterion to model the low-dimensional feature space, which increases robustness to imprecise alignment and strengthens discrimination for facial expressions. Extensive experimental results on Extended Cohn-Kanade and MAHNOB-HCI databases demonstrate that the proposed method achieves the best results in recognizing facial expressions and performs robustly with ordinary on general face detection and eye detection.

Local binary pattern from three orthogonal planes (LBPTOP) has been widely used in emotion recognition in the wild. However, it suffers from illumination and pose changes. We focus the robustness of LBPTOP to unconstrained environments. Our recent proposed method, spatiotemporal local monogenic binary pattern (STLMBP) was verified to work promisingly in different illumination conditions. This improved spatiotemporal descriptor uses not only magnitude and orientation, but also the phase information, which provides complementary information. In detail, the magnitude, orientation and phase images are obtained by using an effective monogenic filter, and multiple feature vectors are finally fused by multiple kernel learning. STLMBP and the proposed method are evaluated in the Acted Facial Expression in the Wild as part of the 2014 Emotion Recognition in the Wild Challenge. They achieve competitive results, with an accuracy gain of 6.35% and 7.65% above the challenge baseline (LBPTOP) over video.

After we started the pioneering work on micro-expression recognition in 2011, this topic has been attracting more and more attention. Spotting micro-expressions is a primary step for continuous emotion recognition from videos. Spotting in this context refers to automatically finding the temporal locations of the face-related events from a video sequence. Rapid facial movements mainly include micro-expressions and eye blinks. However, the role of eye blinks in expressing emotions is still controversial, and often they are considered as micro-expressions as well. In our work a simple method for automatically spotting rapid facial movements from videos is proposed. The method relies on analyzing differences in appearance-based features of sequential frames. In addition to finding the temporal locations, the system is able to provide spatial information about the movements in the face. Micro-expression spotting experiments are carried out on three datasets consisting only of spontaneous micro-expressions. Baseline micro-expression spotting results are provided for these three datasets including the publicly available CASME database. Also an example of spatial localization of the spotted rapid movements is presented.

Heart Rate Measuring from Videos

Remote heart rate (HR) measurement from face videos recorded by ordinary cameras is a new research topic. Previous methods can achieve high accuracies under
well controlled situations, but their performance significantly degrades when environmental illumination variations and subjects’ motions are involved. Our proposed framework contains three major processes to reduce these interferences: first, we employ Discriminative Response Map Fitting method to find the precise face region and use tracking to address the problem caused by rigid head movement; second, Normalized Least Mean Squares adaptive filter is employed to rectify the interferences of illumination variations; third, signal segments with big standard deviation values are discarded in order to reduce the noise caused by sudden non-rigid movements. We have demonstrated that all three processes help to improve the accuracy of HR measurement under realistic Human Computer Interaction (HCI) situations.

Proposed framework for HR measurement from facial videos in realistic HCI situations.

Analysis of Visual Speech

It is known that human speech perception is a bi-modal process which makes use of information not only from what we hear (acoustic) but from what we see (visual). In machine vision, visual speech recognition (VSR), sometimes also referred to as automatic lip-reading is the task of recognizing the utterances through analyzing the visual recordings of a speaker’s talking mouth without any acoustic input. Although visual information cannot in itself provide normal speech intelligibility, it may be sufficient within a particular context when the utterances to be recognized are limited. In such a case, VSR can be used to enhance natural human-computer interactions through speech especially when audio is not accessible or severely corrupted.

We made a detailed review of recent advances in this research area. In comparison with the previous survey which covers the whole ASR system that uses visual speech information, we focus on the important questions asked by researchers and summarize the recent studies that attempt to answer them. In particular, there are three questions related to the extraction of visual features, concerning speaker dependency, pose variation and temporal information, respectively. Another question is about audio-visual speech fusion, considering the dynamic changes of modality reliabilities encountered in practice. In addition, the state of the art on facial landmark localization is briefly introduced. Those advanced techniques can be used to improve the region-of-interest detection, but have been largely ignored when building a visual-based ASR system. We also provide details of audio-visual speech databases. Finally, we discuss the remaining challenges and offer our insights into the future research on visual speech decoding.

Visual speech information plays an important role in automatic speech recognition (ASR) especially when audio is corrupted or even inaccessible. Despite the success of audio-based ASR, the problem of visual speech decoding remains widely open. The key question for us to answer in the problem of VSR is how to characterize the highly dynamic process of uttering in a high-dimensional visual data space. The goal is to learn a compact representation for the visual speech data. We propose a latent variable model to learn the representation. The model is generative in the sense that an observed image sequence is assumed to be generated from one shared latent speaker variable (LSV) and a sequence of latent utterance variables (LUVs). The former accounts for the inter-speaker variations of visual appearances and the latter for the variations caused by uttering. We model the structure of image sequences of the same utterance by a path graph and incorporate the structural information through using the low-dimensional curve embedded within the graph as our prior knowledge on the locations of LUVs in the latent space. In such a way, we can impose soft constraints to penalize values of LUVs that contradict the modelled structure.

Illustration of the generative latent variable model and priors on LUVs. Here $\textbf{l}$ stands for the LSV, $[\textbf{w}_t]$ the LUVs and $[\textbf{x}_t]$ an observed visual speech sequence.

Speech when accompanied by visual component improves speech understanding and enriches the overall perceptual experience. For this reason, Visual Speech Animation (VSA) is an active pursuit for many different contexts in human-computer interaction. Language teaching assistant, virtual avatar, animated story narration are few of those contexts. VSA can be classified as image-based or 3D-shape-based depending on the representation of the renderable visual information. Acquisition of image-based speech corpus is comparatively easier due to the low cost and ease of video recording and its processing. For this reason, image-based VSA systems have been highly successful in producing high quality realistic speech animation.
However, image-based systems are constrained in terms of renderability of face in varying poses, illumination and detail. In contrast, the acquisition of 3D visual speech corpus is computationally expensive and slow, but unlike image-based VSA systems they are flexible in terms of their renderability and detail. Our approach of VSA takes the advantages of both image-based and 3D-shape-based systems by using an already existing image-based system and generating 3D visual speech from its output. We do this by using a small 3D speech corpus when compared to that required for developing a conventional corpus-based 3D VSA system. In fact, this is our prime motivation. Consequently, our system not only has the advantage of the 2D VSA system in producing natural speech dynamics, but also the renderability of a 3D VSA system. The system has two modules, the first module estimates the 3D shape sequence for an input image sequence and, the second module complements the external 3D face with 3D eyes, tongue and teeth.

Visual speech constitutes a large part of our non-rigid facial motion and contains important information that allows machines to interact with human users, for instance, through automatic visual speech recognition (VSR) and speaker verification. One of the major obstacles to research of non-rigid mouth motion analysis is the absence of suitable databases. Those available for public research either lack a sufficient number of speakers or utterances or contain constrained view points, which limits their representativeness and usefulness. We collected a novel multi-view audiovisual database, named OuluVS2, for non-rigid mouth motion analysis. It includes more than 50 speakers uttering three types of utterances and more importantly, thousands of videos simultaneously recorded by six cameras from five different views spanned between the frontal and profile views. The database was preprocessed: videos from different views were synchronized, utterances were located and talking mouth images were extracted.

A simple VSR was developed and tested in an experimental setting that had been widely used in previous VSR studies to provide some baseline performance. Recognition results show that the best VSR performance does not come from the frontal view or those close-to-frontal views. They highlight the need for more research effort to better understand visual speech especially under various camera views.

Head Pose Estimation

Head Pose Estimation (HPE) has recently attracted a lot of interests in various computer vision applications. One challenging problem for accurate HPE is to model the intrinsic variations among poses, and suppress the extraneous variations derived from other factors, such as the illumination changes, outliers, and noise. To this end, we proposed a simple and efficient facial description for head pose estimation from images. To handle the illumination changes, we characterize each image pixel by its image gradient orientation (IGO), rather than the intensity, which is sensitive to illumination changes. We then carry out complex-frequency domain analysis of the IGO image via the two-dimensional image transform, such as the 2D Discrete Cosine Transform (DCT2), to encode the spatial configuration of image gradient orientations. The proposed facial description is called IGO-DCT2. It is robust to illumina-
nation changes, outliers, and noise. In addition, it is learning free and computationally efficient. Finally, the fine-grain head pose estimation is regarded as a regression problem and off-the-shelf non-linear regression models are used to learn the mapping from the feature space to the continuous pose labels. Experimental results show the proposed facial description achieves highly competitive results on the publicly available FacePix dataset.

![Image gradient orientation images under different illumination.](image1)

The image gradient orientation images under different illumination.

![Image IGO images under different poses.](image2)

The IGO images under different poses.

**Affective Human-Computer Interaction**

A paper describing our Minotaurus system developed for affective human-robot interaction in smart environments was published in 2014 in Cognitive Computation journal.

A special face analysis demo was developed for a Mission: Better Life exhibition in Helsinki that showcased different future technologies from Finnish universities as part of Millennium Technology Prize celebrations. This demo showcased various different analysis techniques that are possible to do with face images in real-time using regular cameras. Demo included face recognition, facial expression recognition and gender recognition. Face recognition component is the centerpiece of this analysis as it can dynamically learn new people that appear in front of the camera and remember previous analysis results and combine them with new ones for each different person. It also displays how the recognition is done by visualizing each individual step and their intermediate results. This demo was later extended with a more advanced facial landmark detector and a real-time version of the heart rate measuring system.

![Screenshot of the face analysis demo developed at CMV.](image3)

Screenshot of the face analysis demo developed at CMV.

**Vision Systems Engineering**

Vision systems engineering research aims to identify attractive computing approaches, architectures, and algorithms for industrial machine vision systems. In this research, solutions ranging from low-level image processing even to equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial solutions. We have also applied our expertise to applications intended for embedded mobile platforms.

The key area studied in machine vision based wood inspection was knot detection and accurate localization of its boundary. Knots are one, very common type of defect found in images taken from wood or wooden products. In many cases, the boundary between the knot and the background is not easily distinguishable and traditional methods that rely solely on thresholding tend to exaggerate the knot area. We have developed a method for this problem that can 1) detect possible knot candidates and 2) accurately mark the knot area. The found areas are analyzed further to reveal the actual knot border.

In mobile and embedded platforms, the use reconfigurable computing for vision-based interactive applications and UIs has been studied to identify its trade-offs and challenges. With an emphasis in the impact of the combined consideration of computing, sensing and interactivity, three reconfigurable architectures, an EnCore processor with a Configurable Flow Accelerator, a hybrid SIMD/MIMD reconfigurable coprocessor, and Transport-Triggered Architecture processors, have been analyzed in terms of performance and energy efficiency. The advantages of integrating dedicated reconfigurable resources in mobile devices have an
impact on the adaptation of the design principles at a platform level.

On mobile graphics processors, the identification of missing and unsupported abstractions of the current mobile graphics processing units APIs and tool-chains has brought vision-based interactive computing to other developers. This provides for novel insight into efficient high-performance mobile GPGPU implementation of interactive applications.

Accurate detection of knot boundaries.

We have also continued the research on multimodal gesture controlled user interaction. The methods developed work with the already existing hardware in recent mobile devices. The gestures are recognized from the front camera and the touch screen. With the user interface, the user can move the mouse cursor, click on objects and scroll documents. The functions provided to the user depend on the distance between the hand and the device. For this purpose, we have continued to develop our finger tracking and detection systems by testing the use of some widely used features in our system.

The evolution of mobile and embedded systems.

In the energy efficient architectures and signal processing area, we have been working on design automation and energy efficient computing for signal processing applications. The joint US-Finnish research project "CREAM", together with the Centre for Wireless Communications, has provided publications related to dataflow modeling and energy-efficient implementation of a digital predistortion filter for wireless mobile transmitters. In the context of video processing, a programmable, energy-efficient multicore processor for HEVC/H.265 joint deblocking- and sample adaptive offset filtering was developed. A new opening has been starting research collaboration with the team of Prof. Lothar Thiele of ETH Zürich.

Programmable parallel accelerator for multimedia applications design work continued. Two computationally intensive algorithms; face detection and depth estimation, were implemented and optimized for parallel processing using the Portable computing language (PoCL) implementation of Open Computing Language (OpenCL). The accelerator is being benchmarked against desktop and mobile GPUs for performance and energy efficiency comparisons. In addition, research aiming for heterogeneous accelerator for both multimedia and wireless communication was started.

Biomedical Image Analysis

In recent years, increasing resolving power and automation of biomedical imaging systems have resulted in an exponential growth of the image data. Manual analysis of these data sets is extremely labor intensive and hampers the objectivity and reproducibility of results. Hence, there is a growing need for automatic image processing and analysis methods. In CMV, our aim has been to apply modern computer vision techniques to biomedical image analysis which is one of our emerging research areas.

As a part of our ABCdata project funded by TEKES, we proposed a novel local image descriptor that can be utilized in various applications in biomedical image analysis research. Our proposed feature descriptor is based on statistical models of images. In this work, performance of the descriptor is tested in a microscopy image pixel labelling framework. Pixel level identification scheme can be employed as a generic detection method and as a priori for subsequent segmentations of different cell lines and microscopy modalities. We employ our feature descriptor for detection of tumor cell spheroids in phase contrast imaging of cell cultures and for detection of mitochondria in electron microscopy images. Our method works under heavy occlusions and clutter and therefore suitable for most of the biomedical images. Experimental results demonstrate significant improvements over a strong baseline Scale Invariant Feature Transform (SIFT) descriptor.

Proposed pixel classification pipeline.

In our research related to biomedical image analysis research, we have also studied cell morphology. Automated analysis of cell morphology is important in order to have an understanding of the relationship between
cell shape and cell culture. In this context we analyzed branching characteristics of cell clusters over time in a quantitative manner. We started with the binary image resulting from the cell segmentation process. Morphology of each cell cluster is then analyzed individually by employing contour convexity defects. We associate the number of defects with the number of branches and the distance between the farthest contour point and the hull is utilized to account for cell extension (branch) size.

**Illustration of our cell morphology analysis.**

Imaging of living cells is becoming a key tool to answer the fundamental questions about cell dynamics, molecular regulation of cell migration, cell invasion and cell fates. 3D scaffolds provide more realistic platform for cell cultures with respect to the physical and biochemical properties of the micro-environment compared to 2D models. 3D scaffolds are especially important in tumor cell cultures in which the composition of the micro-environment contributes to the cell behavior and drug response. These experiments generate huge datasets, which are often infeasible to analyze manually so a good cell segmentation is crucial. This task can be very challenging due to many factors including irregular cell shapes, non-uniform intensities, frequent cell-cell contacts, and presence of other structures.

We have developed a cell segmentation method which can handle complex flexible cell shapes as it does not make any assumption about cell shapes. It utilizes the edge probability map and graph cuts to find seeds for individual cells. Edge probability is computed at each pixel using information in its local neighborhood. Then it finds seeds for cells by finding min-cut solution of the grid graph, whose terminal edge weights are set using edge probability map and pixel intensities in local neighborhood of a pixel. Then we use Marker-controlled watershed to expand the seeds and find segmentation.

**Overview of our cell segmentation method (above), and the maximum intensity projection of a 3D stack and its segmentation where cell are labelled with colors (below).**

In 2014, CMV and OEM have been working on an ITEE collaboration project which consists of both hardware and software components. Optical tweezers (OT) is a novel tool that allows for noncontact trapping and manipulating single micro- and nanosized particles using tightly focused laser beam. The most perspective feature of OT is that they can measure forces ranging from several pN to almost hundred pN. Such forces characterize the interactions between biological cells and macromolecules.

In order to make quantitative measurements, OT must be calibrated. Shortly speaking, calibration for optical tweezers means knowing what the trap stiffness and trap force for a laser beam are at certain intensities. Consequently, the force experienced by an object as a function of a measured displacement of the object in the trap is then known. The power spectrum analysis of the Brownian motion of a trapped particle is usually considered to be the most reliable option to accomplish this goal. Different methods, such as quadrant photo detector (QPD) or fast video camera can be used for detecting the particle during the calibration. During 2014, we have develop further the calibration method based on video recordings. Since the typical amplitude of the motion of a trapped particle is of the order of tens of nanometers, position determination from the captured images must be carried out with sub-pixel accuracy.
Osteoarthritis (OA) causes progressive degeneration of articular cartilage and pathological changes in subchondral bone. These changes can be assessed volumetrically using micro-computed tomography (µCT) imaging. The local descriptor, i.e. local binary pattern (LBP), is a new alternative solution to perform analysis of local bone structures from µCT scans. In this study, trabecular bone samples were prepared from patients treated with total knee arthroplasty and the LBP descriptor was applied to correlate the distribution of local patterns with the severity of the disease. The results obtained suggest the apparition and disparition of specific oriented patterns with OA, as an adaptation of the bone to the decrease of cartilage thickness. The experimental results suggest that the LBP descriptor can be used to assess the changes in the trabecular bone due to OA.

Exploitation of Results

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern methodology and its variants are used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis. The researchers in CMV have actively published the source codes of their algorithms for the research community, and this has increased the exploitation of the results.

The results have also been utilized in our own projects. For example, we have collaborated with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with physiological biosignals. Together with Prof. Seppänen and Dr. Seppo Laukka (Department of Educational Sciences and Teacher Education) and Prof. Matti Lehtihalms (Faculty of Humanities) we have participated in the FSR Second Wave project where we have developed a Mobile Multimodal Recording System (MORE) that is now actively used in classroom research in various schools. With Assoc. Prof. Simo Saarakkala (Faculty of Medicine) we have investigated LBP-based methodology for the diagnosis of osteoarthritis.

Most of our funding for both basic and applied research comes from public sources such as the Academy of Finland and Tekes, but besides these sources, CMV also conducts research by contract which is funded by companies. In this way, our expertise is being utilized by industry for commercial purposes, and even in consumer products, like mobile devices.

The CMV has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience, their roots are especially in the strands of “free academic research”. There are currently altogether five research based spin-outs founded directly on the machine vision area. The number of spin-outs could be extended up to sixteen when taking into account the influence of the CMV’s over thirty-year old history and the spin-out companies from the spin-out research groups in the area of computer science and engineering in total.

Future Goals

The very positive results obtained, e.g. from the RAE 2014 evaluations show that we are on the right track. We plan to carry out well focused cutting-edge research, for example, on novel image and video descriptors, perceptual interfaces for face to face interaction, multimodal analysis of emotions, 3D computer vision, biomedical image analysis, and energy-efficient architectures for embedded vision systems. We also have plans to further deepen our collaboration with international and domestic partners. For this purpose, we are participating in new European project proposals. We are also active in applying funding for breakthrough research from the European Research Council (ERC), obtaining recently very promising evaluation results. Close interaction between basic and applied research has always been a major strength of our research unit. The scientific output of the CMV has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.
Selected Publications


IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Florence, Italy, 8336-8340.


He Q, Hong X, Zhao G & Huang X (2014) An immersive fire training system using Kinect. UbiComp 2014 Adjunct Proceedings, Seattle, USA, accepted.


Background and Mission

The Center for Machine Vision Research (CMV) is a creative, open and internationally attractive research unit. It is renowned world-wide for its expertise in computer vision, which now spans for nearly 35 years.

The Center has achieved ground-breaking research results in many areas of its activity, including texture analysis, facial image analysis, 3D computer vision, and energy-efficient architectures for embedded systems. Among the highlights of its research are the Local Binary Pattern (LBP) methodology, LBP-based face descriptors, and methods for geometric camera calibration, which all are highly-cited and widely used around the world. The areas of application for CMV’s current research include affective computing, perceptual interfaces for human-computer interaction, biometrics, augmented reality, and biomedical image analysis. The CMV has a wide international collaboration network to support its research mobility.

In spring 2016, the staff of CMV consists of three Professors, one Associate Professor, two FiDiPro Professors and one FiDiPro Fellow, 15 senior or postdoctoral researchers, and 16 doctoral students or research assistants. We have also visiting scholars coming with their own funding. The unit is highly international: over 50% of our researchers (doctors, PhD students) come from abroad. CMV has an extensive international collaboration network in Europe, China, Japan, Australia, and the USA. Both outgoing and incoming mobility of researchers is intense to/from leading research groups abroad. In 2015, CMV participated in two European COST actions.

Scientific Progress

The current main areas of research are: 1) Computer vision methods, 2) Human-centered vision systems, 3) Vision systems engineering, and 4) Biomedical image analysis.

The group has a long and highly successful research tradition in two important generic areas of computer vision: texture analysis and geometric computer vision. Vision systems engineering has been a basis for many practical machine vision applications developed in the group. In recent years, computational photography, object detection and recognition, and biomedical image analysis have also become important research topics in CMV. The aim in all these areas is to create a methodological foundation for the development of new vision-based technologies and innovations.

Highlights and Events in 2015

In 2015, CMV received funding for two FiDiPro projects from Tekes – The Finnish Funding Agency for Innovation. Two distinguished computer vision scientists, Jiri Matas (Czech Technical University in Prague) and Stefanos Zafeiriou (Imperial College London), will visit the University of Oulu on a regular basis in 2016-2019 and contribute to a joint research agenda. Both FiDiPro projects will also benefit a set of companies as project partners.

Professor Xilin Chen, whose term as FiDiPro Professor will end in 2016, was elevated to IEEE Fellow. He is being recognized for his contributions to machine vision for facial image analysis and sign language recognition.

According to a Web of Science analysis, CMV continues to excel in producing highly-cited papers, which is a commonly used measure for breakthrough research. In 2005-15, it had five papers in top-20, among all 18,522 Finnish papers (19.4.2016) published in Engineering category of Web of Science during that period, ranking 2nd, 5th, 7th, 8th, and 20th.

In November, our research on hidden facial expressions was reported by MIT Technology Review. The article was based on our report on reading hidden emotions from spontaneous micro-expressions published in arXiv.

In May, the 2015 MVA Most Influential Paper over the Decade Award was granted to Dr. Vili Kellokumpu, Prof. Matti Pietikäinen and Prof. Janne Heikkinä. This award is given to the authors of papers that were presented at the IAPR International Conference on Machine Vision Applications (MVA) held ten years before (this time MVA 2005) and have been recognized as having the most significant influence on machine vision technologies over the subsequent decade.

CMV members have been active in co-editing special issues in prestigious journals: Dr. J. Chen, Dr. G. Zhao and Prof. M. Pietikäinen, together with Dr. Z. Lei (from Chinese Academy of Sciences) and Dr. L. Liu (from National University of Defense Technology, China) co-edited the Special Issue on robust local de-
criptors for computer vision in Neurocomputing journal; Dr. G. Zhao and Prof. M. Pietikäinen, together with S. Zafeiriou, I. Kotsia (from Middlesex University, UK), J. Cohn (from University of Pittsburgh/CMU, USA) and R. Chellappa (from University of Maryland), have been co-editing the Special issue on Spon-
taneous Facial Behaviour Analysis (SFBA) for Com-
puter Vision and Image Understanding journal; Prof.
Heikilä together with Prof. L. Xie and Dr. Zhang
(both from Northwestern Polytechnic University, Chi-
a) co-edited the Special Issue on Immersive Au-
dio/Visual Systems in Multimedia Tools and Appli-
cations journal.

Computer Vision Methods

Texture Analysis

Texture is an important characteristic of many types of
images and can play a key role in a wide variety of
applications of computer vision and image analysis.
The CMV has long traditions in texture analysis re-
search, and ranks among the world leaders in this area.
The Local Binary Pattern (LBP) texture operator has
been highly successful in numerous applications
around the world, and has inspired plenty of new re-
search on related methods, including the blur-
sensitive Local Phase Quantization (LPQ) method,
Weber Law Descriptor (WLD), and Binarized Statisti-
cal Image Features (BSIF), also developed by research-
ers of CMV.

We proposed a globally rotation invariant multi-scale
cocurrence local binary pattern (MCLBP) feature for
texture-relevant tasks. In MCLBP, we arrange all
cocurrence patterns into groups according to proper-
ties of the co-patterns, and design three encoding func-
tions (Sum, Moment, and Fourier Pooling) to extract
features from each group. The MCLBP can effectively
capture the correlation information between different
scales and is also globally rotation invariant (GRI). The
MCLBP is substantially different from most existing
LBP variants including the LBP, the CLBP, and the
MSJ-LBP that achieves rotation invariance by locally
rotation invariant (LRI) encoding. Extensive experi-
ments demonstrated the effectiveness of the MCLBP
compared to many state-of-the-art LBP variants includ-
ing the CLBP and the LBPHF.

We also proposed a local feature, called Local Orien-
tation Adaptive Descriptor (LOAD), to capture regional
texture in an image. In LOAD, we propose to define
point description on an Adaptive Coordinate System
(ACS), adopt a binary sequence descriptor to capture
relationships between one point and its neighbors and
use multi-scale strategy to enhance the discriminative
power of the descriptor. The proposed LOAD enjoys
not only discriminative power to capture the texture
information, but also has strong robustness to illumina-
tion variation and image rotation. Extensive experi-
ments on benchmark data sets of texture classification
and real-world material recognition show that the
LOAD yields the state-of-the-art performance. By
combining LOAD with Convolutional Neural Net-
works (CNN), we obtain significantly better perform-
ance than both the LOAD and CNN. This result con-
firms that the LOAD is complementary to the learning-
based features.

LBPs are considered among the most computationally
efficient high-performance texture features. However,
the LBP method is very sensitive to image noise and is
unable to capture macrostructure information. To best
address these disadvantages, we collaborated with Dr.
Li Liu, (National University of Defense Technology,
China) and others, to introduce a novel descriptor for
texture classification, the Median Robust Extended
Local Binary Pattern (MRELBP). Different from traditional
LBP and many LBP variants, MRELBP com-
pares regional image medians rather than raw image
intensities. A multiscale LBP type descriptor is com-
piled by efficiently comparing image medians over a
novel sampling scheme, which can capture both micro-
structure and macrostructure texture information. A
comprehensive evaluation on benchmark datasets re-
veals MRELBP’s high performance - robust to gray
scale variations, rotation changes and noise - but at a
low computational cost. MRELBP has produced the
best classification scores on many different test data-
bases. More importantly, MRELBP is highly robust to
image noise including Gaussian noise, Gaussian blur,
Salt-and-Pepper noise and random pixel corruption. A
paper on this method was published in IEEE Transac-
tions on Image Processing.

Dynamic texture and scene classification are two funda-
mental problems in understanding natural video
content. Extracting robust and effective features is a
crucial step towards solving these problems. However,
the existing approaches suffer from the sensitivity to
either varying illumination, or viewpoint changes, or
even camera motion, and/or the lack of spatial infor-
mation. Inspired by the success of deep structures in
image classification, we attempted to leverage a deep
structure to extract features for dynamic texture and
scene classification. To tackle with the challenges in
training a deep structure, we propose to transfer some
prior knowledge from image domain to video domain.
To be more specific, we propose to apply a well-
trained Convolutional Neural Network (ConvNet) as a
feature extractor to extract mid-level features from
each frame, and then form the video-level representa-
tion by concatenating the first and the second order
statistics over the mid-level features. We term this two-
level feature extraction scheme as a Transferred Con-

VNet Feature (TCoF). Moreover, we explore two dif-
ferent implementations of the TCoF scheme, i.e., the
spatial TCoF and the temporal TCoF. In the spatial
TCoF, the mean-removed frames are used as the inputs
of the ConvNet; whereas in the temporal TCoF, the
differences between two adjacent frames are used as
the inputs of the ConvNet. We evaluated systematically
the proposed schemes on three benchmark data sets,
including DynTex, YUPENN, and Maryland, demon-
stratifying that the proposed approach yields excellent performance.

Local descriptors are popular ways to characterize the local properties of images in various computer vision based tasks. To form the global descriptors for the image regions, the first-order feature pooling is widely used. However, as the first-order pooling technique treats each dimension of local features separately, the pairwise correlations of local features are usually ignored. Encouraged by the success of recently developed second-order pooling techniques, we formulate a general second-order pooling framework and explore several analogues of the second-order average and max operations. We comprehensively investigate a variety of moments which are in the central positions to the second-order pooling technique. As a result, the superiority of the second-order standardized moment average pooling (2Standmap) is suggested. The 2Standmap provides a unified approach to capsule both low-level information from raw features and the mid-level visual cues from the local descriptors. It is of low dimension, discriminative, efficient, and learning free. We successfully apply 2Standmap to four challenging tasks namely texture classification, medical image analysis, pain expression recognition, and micro-expression recognition. It illustrates the effectiveness of 2Standmap to capture multiple cues and the generalization to both static images and spatial-temporal sequences.

Illustration of proposed background subtraction framework, and results of detected foreground and recovered background.

Low rank and sparse representation based methods, which make few specific assumptions about the background, have recently attracted wide attention in background modeling. With these methods, moving objects in the scene are modeled as pixel-wised sparse outliers. However, in many practical scenarios, the distributions of these moving parts are not truly pixel-wised sparse but structurally sparse. Meanwhile a robust analysis mechanism is required to handle background regions or foreground movements with varying scales. Based on these two observations, we first introduce a class of structured sparsity-inducing norms to model moving objects in videos. In our approach, we regard the observed sequence as being constituted of two terms, a low-rank matrix (background) and a structured sparse outlier matrix (foreground). Next, in virtue of adaptive parameters for dynamic videos, we propose a saliency measurement to dynamically estimate the support of the foreground. Experiments on challenging well known data-sets demonstrate that the proposed approach outperforms the state-of-the-art methods and works effectively on a wide range of complex videos.

Background Subtraction

Foreground object segmentation from a video stream is a fundamental and critical step for many high level computer vision tasks. The accuracy of segmentation can significantly affect the overall performance of the application employing it. Background subtraction is generally regarded as an effective method for extracting the foreground. However, the background in a complex environment may include distracting motions and hence makes precise segmentation challenging.

Overview of the second pooling framework 2Standmap. Given an image region I, for each pixel x inside, a vector f(x) concatenating all local features is extracted. We then perform the second order pooling to obtain the global region descriptor g. Two key parts of the second order pooling namely 2nd order average or max pooling, and Non Linear mapping are comprehensively investigated in our work.

Computational Photography

In computational photography, the aim is to develop techniques for computational cameras that give more flexibility to image acquisition and enable more advanced features to be employed, going beyond the capabilities of traditional photography. These techniques often involve use of special optics and digital image processing algorithms that are designed to eliminate the degradations caused by the optical system and viewing conditions.

Multi-aperture camera refers to an imaging device that consists of several camera units, each having dedicated optics and color filter. The camera produces several sub-images, which are combined into a single RGB image. Such camera is a feasible alternative to traditional Bayer filter camera in terms of image quality, camera size and camera features.
The main challenge of the multi-aperture camera arises from the fact that each camera unit has a slightly different viewpoint. We have developed a novel image fusion algorithm that corrects the parallax error between the sub-images using a disparity map. In order to improve the disparity estimation we combine the matching costs over multiple views. Promising test results imply that multi-aperture camera has potential to become a serious competitor to Bayer filter cameras in portable devices.

Object Detection and Recognition

Nowadays object detection and recognition is an important research area in computer vision as it has potential to facilitate search from very large unannotated image databases. However, even the best current systems for artificial vision have a very limited understanding of objects and scenes. For instance, state-of-the-art object detectors model objects as distributions of simple features (e.g., HOG or SIFT), which capture a blurred statistics of the two-dimensional shape of the objects. Color, material, texture, and most of the other object attributes are likely ignored in the process.

Automatic Video Content Annotation

TV and movie industry has seen a paradigm shift during the last years, as the Internet has become one of the most important distribution channels. Platforms, such as the BBC’s iPlayer, and distributors and portals like Netflix, Amazon Instant Video, Hulu, and Youtube, have millions of users every day. Due the amount of media, the search is essential functionality in all platforms. These are typically based on video title and metadata, but not on actual visual context. Enabling such search option would require annotating the video content, which is very laborious to do manually. In this project we aim to enable fully automatic system that labels video based on information fusion from the visual and textual sources.

Most often the video is accompanied textual information in the form of subtitles and transcripts. Subtitles contain the information on who is saying what and when, whereas transcripts describe actions in the scene on a higher level without very precise timing. In our research, we have developed a learning system that is able to utilize this kind of weak supervision to learn to detect and recognize the actors appearing in the movies. Moreover, unlike previous works, we are able to detect and classify characters who are not speaking at all.

One of the very recent innovations is a system that utilizes higher level cues like timely overlapping face tracks and overall number of detections per character in the decision making process. In addition, we learn novel convolutional neural network features that represent entire face tracks. These features provide to be superior compared to previous generation methods based on hand-crafted descriptors or individual frame representations.

Learning Based Image Representations

Finding correspondences between image regions (patches) is a key factor in many computer vision applications. Structure-from-motion, multi-view reconstruction, image retrieval and object recognition require an accurate computation of similarity. Due to importance of these problems various descriptors have been proposed for patch matching with the aim of improving accuracy and robustness. Many of the most
widely used approaches, like SIFT or DAISY descriptors, are based on hand-crafted features and have limited ability to cope with negative factors (occlusions, variation in viewpoint etc.) making a search of similar patches more difficult. Recently, various methods based on supervised machine learning have been successfully applied to learning patch descriptors.

In our research, we have developed a new learning based image patch descriptor. The new descriptor is learned directly from the data using a Siamese convolutional neural network constructions. A key feature in the obtained descriptor is that by design the pairwise comparisons are possible using simple Euclidean distance. This ability, which is not true for all learning based descriptors, make the utilization of this work simpler.

Semantic illustration of the Siamese network structure that is utilized to learn efficient image patch descriptors.

3D Computer Vision

Three-dimensional scene reconstruction from multiple images has been a popular research topic in computer vision for several decades. During recent years, there has been significant progress both in large-scale structure from motion (SfM) and simultaneous localization and mapping (SLAM). One example is our DT-SLAM system published in late 2014, which is an open source implementation for real-time SLAM.

In 2016, we have focused on 3D modeling from point clouds produced with image-based 3D reconstruction techniques. In particular, we developed a method for evaluating triangle mesh models. The method takes the reconstructed mesh or point cloud and the ground truth as input and outputs two values, namely Jaccard index and compression ratio, which represent the accuracy, completeness and compactness of the reconstruction. The previous evaluation metrics usually measure only the accuracy and completeness of the mesh whereas our method is able to measure the compactness-accuracy trade-off of the reconstructions. The evaluation method presented in the figure below.

Range cameras have greatly simplified many 3D vision problems that have been considered to be extremely difficult using conventional imaging techniques. While Microsoft Kinect was the first inexpensive range sensor that gained a lot of interest in the computer vision community, range cameras are now also emerging to mobile devices. Google Tango has been the first mobile platform that embeds many sensors including a range camera. It provides both depth maps and color images (so called RGB-D data) and also relatively good and robust odometry, which is computed by combining visual and inertial sensor information. Our aim has been to correct the errors, e.g. drift, that the Tango’s built-in hardware and software produce, and thereby improve the final outcome of the 3D scene reconstruction.
addition, we introduced a publicly available range object database which is large and has a high diversity that is suitable for similarity retrieval applications. The simulation results indicate competitive performance between local and global methods. While better complexity trade-off can be achieved with the global techniques, local methods perform better in distinguishing different parts of incomplete depth data. It could also be concluded that global and local descriptors can be merged to achieve a higher performance in depth data similarity retrieval applications.

Human-Centered Vision Systems

In future ubiquitous environments, computing will move into the background, being omnipresent and invisible to the user. This will also lead to a paradigm shift in human-computer interaction (HCI) from traditional computer-centered to human-centered systems. We expect that computer vision will play a key role in such intelligent systems, enabling, for example, natural human-computer interaction, or identifying humans and their behavior in smart environments.

Face Recognition and Biometrics

We continued our research on unconstrained face recognition by investigating the recent binarized statistical image features (BSIF) local descriptor method. BSIF is a learning based method that can be used to learn application specific descriptors. In our first work, we combined BSIF and a novel soft-assignment method yielding state-of-the-art results with the well-known FERET database. The proposed method performed well also in unconstrained settings where we evaluated its performance on the LFW database. Motivated by the positive results using BSIF, we performed a review of unsupervised feature learning methods and found a method called Reconstruction ICA (RICA) which is quite similar to the ideas behind BSIF. Using RICA and the popular Bag-of-Features model we constructed a face representation method that was able to outperform the state-of-the-art in the unsupervised evaluation category of the original LFW protocol.

We also continued our research on face biometrics under spoofing attacks. While previous research has mainly been focused on analyzing the luminance of the face images, hence discarding the chrominance information which can be useful for discriminating fake faces from genuine ones, we proposed a new face anti-spoofing method based on color texture analysis. We analyzed the joint color-texture information from the luminance and the chrominance channels using a color local binary pattern descriptor. More specifically, the feature histograms are extracted from each image band separately. Extensive experiments on two benchmark datasets showed excellent results compared to the state-of-the-art. This work was ranked among the best 10% papers presented and published at the IEEE International Conference on Image Processing (ICIP 2015).

Continuing the anti-spoofing research theme, we investigated audiovisual synchrony assessment for replay attack detection in talking face biometrics. Audiovisual speech synchrony detection is an important liveness check for talking face verification systems in order to make sure that the input biometric samples are actually acquired from the same source. In prior work, the used visual speech features have been mainly describing facial appearance or mouth shape in frame-wise manner, thus ignoring the lip motion between consecutive frames. Since also the visual speech dynamics are important, we took the spatiotemporal information into account and proposed the use of space-time auto-correlation of gradients (STACOG) for measuring the audiovisual synchrony. For evaluating the effectiveness of the proposed approach, a set of challenging and realistic attack scenarios were designed. Extensive experimental analysis showed that the STACOG features outperform the state of the art in measuring the audiovisual synchrony.

Although still-to-video face recognition is an important function in watch list screening, state-of-the-art systems often yield limited performance due to camera inter-operability and to variations in capturing conditions. Therefore, the visual comparison of faces captured in unconstrained low-quality videos against a matching high-quality reference facial still image cap-
ured under controlled conditions is required in many surveillance applications. To improve the visual appearance of faces captured in videos, we studied a new super-resolution (SR) pipeline that is suitable for fast adjudication of face-matches produced by an automated system. In this pipeline, face quality measures are used to rank and select face captures belonging to a facial trajectory, and multi-image SR iteratively enhances the appearance of a super-resolved face image. Face selection is optimized and registered using graphical models. Experimental results showed that the proposed pipeline efficiently produces super-resolved face images by ranking best quality ROIs in a trajectory. To select the best face captures for SR, this pipeline exploits a strong correlation existing between pose and sharpness quality measurements. This work was done in a close collaboration with Ecole de Technologie Superieure, Universite du Quebec, Canada.

We co-organized a special issue on the theme of soft biometrics at the Pattern Recognition Letters. Some of the articles are extended versions of selected papers presented at the "1st International Workshop on Soft Biometrics," which was held in conjunction with the European Computer Vision Conference (ECCV 2014) in Zurich on September 7th 2014 and promoted by the European COST Action IC1106 "Integrating Biometrics and Forensics for the Digital Age." The special issue featured a review paper entitled "On Soft Biometrics," written by experts in soft biometrics research. The article provided an introduction to the topic and an overview of the history and achievements in the extraction and use of soft biometrics traits. A new definition of the term is also presented: "the estimation or use of personal characteristics describable by humans that can be used to aid or effect person recognition." The other articles covered a wide range of aspects of soft biometrics research and provide an excellent summary of the state-of-the-art and recent developments in this evolving field. The articles included in this special issue are just samples of work being carried out by different research groups that illustrate some of the soft biometrics traits, their potential applications, and many open issues and challenges. We believe the articles provide important insights into this relatively new research field. We hope that the readers enjoy reading these selected articles and find novel research avenues and agendas that can push forward the field of soft biometrics. The guest editors of this special issue were Paulo Lobato Correia (Portugal), Abdenour Hadid (Finland), Thomas B. Moeslund (Denmark).

How much information Microsoft Kinect facial depth data can reveal about identity, gender and ethnicity? To answer this research question, we explored the usefulness of the depth images provided by Microsoft Kinect sensors in different face analysis tasks including identity, gender and ethnicity. Four local feature extraction methods (LBP, LPQ, HOG and BSIF) were investigated for both face texture and shape description. Extensive experimental analysis on three publicly available Kinect face databases was conducted. The experimental analysis yields into interesting findings. Furthermore, a comprehensive review of the literature on the use of Kinect depth data in face analysis was provided along with the description of the available databases.

We investigated kinship verification from faces using different local texture features obtaining promising results. During the experimental analysis, we noticed that many of the existing kinship databases are limited or biased. For instance, the Kinship Face in the Wild data sets (KinFaceW I & II), published in TPAMI, are commonly used for the evaluation of kinship verification algorithms. We noted that the images in these data sets have relationship pairs that in many cases have been cropped from the same original images. This cropping fact, when taken into account, can significantly bias and simplify the classification problem. A classification strategy that tries to determine if both images in a pair are cropped from the same photo will show improvements when compared to approaches focusing only on facial features. To illustrate this anomaly, we presented an extremely simple classification method that requires no training and offers comparable results to the ones obtained with sophisticated methods under the same experimental protocol. This calls the research community for joint efforts to design new and more reliable databases and evaluation protocols to advance the kinship verification research. For fair comparison, the new databases should include the conditions of the image acquisition and discuss the potential implication of the limitations of the data sets. To ensure the reproducibility of the results, the source code of the different methods should be made publicly available.

**Recognition of Facial Expressions and Emotions**

The face is the key component in understanding the emotions, and this plays significant roles in many areas, from security and entertainment to psychology and education. Micro-expressions (MEs) are rapid, involuntary facial expressions which reveal emotions that people do not intend to show. Studying MEs is valuable as recognizing them has many important applications, particularly in forensic science and psychotherapy. However, analyzing spontaneous MEs is very challenging due to
their short duration and low intensity. In particular, spontaneous MEs (in contrast to posed MEs) are highly challenging due to their large variability in both appearance and duration. Recently, there are increasing interests in inferring micro-expression from facial image sequences. Automated computer vision analysis of micro-expressions consists of emotion recognition and detection in the video.

For micro-expression recognition, feature extraction is an important critical issue. In our research, we proposed two new spatiotemporal feature descriptors for analyzing micro-expression.

In the first work, we proposed a novel framework based on a new spatiotemporal facial representation (called as spatiotemporal local binary pattern with integral projection) to analyze micro-expressions with subtle facial movement. Firstly, we propose to use an integral projection method based on difference images for obtaining horizontal and vertical projection, which can preserve the shape attribute of facial images and increase the discrimination for micro-expressions. Furthermore, we employ the local binary pattern operators to extract the appearance and motion features on horizontal and vertical projections. Intensive experiments are conducted on three publicly published micro-expression databases for evaluating the performance of the method. Experimental results demonstrate that the new spatiotemporal descriptor can achieve promising performance in micro-expression recognition.

Micro-expression detection problem is defined as locating the onset frame (when the facial muscles start moving), peak frame (when the muscle achieves maximum contraction) and offset frame (when the movement disappears). Currently, all the research work on spontaneous micro-expression detection has focused on detecting the peak frames. We were motivated to work on detecting not only the peak frame but also the onset and offset frames. So, we decided to explore motion features because they capture the direction of movements of the face at every frame instance, as illustrated in images (a)-(c) and the trajectory changes at FACS Action Units (as shown in image (d)) can be used to find peak, onset and offset; the motion vectors from onset to peak follows a smooth trajectory and traverses this trajectory in opposite direction from peak to offset. In this approach, the optical flow vectors are added for each AU group across time and their magnitude is plotted as shown in image (e). In this magnitude plot, the peak frame corresponds to the peak and the appropriate onset and offset frames are searched. The drawback is that previous head motions, facial expressions and micro-expressions are accumulated. In order to tackle this problem, the optical flow vector is added across variable small time windows. Various heuristics are used to filter false detections. Experimental results on SMIC spontaneous micro-expression dataset demonstrate that our framework achieves more accuracy than our previous baseline work.
Group Emotion Analysis

Social media has provided much opportunity for people to socially engage and interact with a larger population. In recent years, millions of images and videos have been uploaded on the Internet (e.g. in YouTube and Flickr), enabling us to explore images from a social event, such as family party. However, until recently, relatively little research has examined group emotion in an image. To advance affective computing research, it is indeed of interest to understand and model the affect exhibited by a group of people in images. As we know, feature extraction and group expression model are two critical issues to infer the emotion of a group. In our work, we propose a new method to estimate happiness intensity of a group of people in an image. Firstly, we combine Riesz transform and local binary pattern descriptor, namely Riesz-based volume local binary pattern, which considers neighboring changes not only in the spatial domain of a face but also along the different Riesz faces. Secondly, we exploit to use continuous conditional random fields for constructing a new group expression model, in which considers global and local attributes. Finally, we utilize this model based on Riesz-based volume local binary pattern for estimating group happiness intensity. Numerous experiments are performed on three challenging facial expression databases to evaluate the novel feature. Furthermore, experiments are conducted on HAPPEI database to evaluate group expression model with the new feature. Our experiment results demonstrate our method can provide considerable performance for group happiness intensity analysis.

Multi-Modal Emotion Recognition

Automatic analysis of human spontaneous behavior has attracted increasing attention in recent years from researchers in computer vision. This paper proposes an approach for multi-modal video-induced emotion recognition, based on facial expression and electroencephalogram (EEG) technologies. Spontaneous facial expression is utilized as an external channel. A new feature, formed by percentage of nine facial expressions, is proposed for analyzing the valence and arousal classes. Furthermore, EEG is used as an internal channel supplementing facial expressions for more reliable emotion recognition. Discriminative spectral power and spectral power difference features are exploited for EEG analysis. Finally, these two channels are fused on feature-level and decision-level for multi-modal emotion recognition. Experiments are conducted on MAHNOB-HCI database, including 522 spontaneous facial expression videos and EEG signals from 27 participants. Moreover, human perception in emotion recognition compared to the proposed approach is tested with ten volunteers. The experimental results and comparisons with the average human performance show the effectiveness of the proposed multi-modal approach.

Heart Rate Measuring from Videos

Remote heart rate (HR) measurement from face videos recorded by cameras is a new research topic.

Intel’s newly-announced low-cost and high precision RealSense 3D (RGBD) camera is becoming ubiquitous in laptops and mobile devices starting this year, opening the door for new applications in the mobile health
arena. We demonstrate how the Intel RealSense 3D camera can be used for low-cost gaze tracking and passive pulse rate estimation. We develop a novel 3D gaze and fixation tracker based on the eye surface geometry as well as an illumination invariant pulse rate estimation method using near-infrared images captured with RealSense. We achieve a mean error of 1 cm at 20 × 30 cm for the gaze tracker and 2:26 bpm (beats per minute) for pulse estimation, which is adequate in many medical applications, demonstrating the great potential of novel consumer-grade RGBD technology in mobile health.

Cheek region segmentation using infrared and depth image. (a) Facial landmarks tracked on infrared image. (b) Connected components in depth image. (c) The connected component containing the most landmarks is selected as the face. (d) Face region in infrared image. (e) Cheek area is selected as the region between the eyes and mouth landmarks.

Heart rate measure for ROI in infrared channel.

Analysis of Visual Speech

It is known that human speech perception is a bi-modal process which makes use of information not only from what we hear (acoustic) but from what we see (visual). In machine vision, visual speech recognition (VSR), sometimes also referred to as automatic lip-reading is the task of recognizing the utterances through analyzing the visual recordings of a speaker’s talking mouth without any acoustic input. Although visual information cannot in itself provide normal speech intelligibility, it may be sufficient within a particular context when the utterances to be recognized are limited. In such a case, VSR can be used to enhance natural human-computer interactions through speech especially when audio is not accessible or severely corrupted.

Visual speech constitutes a large part of our nonrigid facial motion and contains important information that allows machines to interact with human users, for instance, through automatic visual speech recognition (VSR) and speaker verification. One of the major obstacles to research of non-rigid mouth motion analysis is the absence of suitable databases. Those available for public research either lack a sufficient number of speakers or utterances or contain constrained view points, which limits their representativeness and usefulness. We introduced a newly collected multi-view audiovisual database for non-rigid mouth motion analysis. It includes more than 50 speakers uttering three types of utterances and more importantly, thousands of videos simultaneously recorded by six cameras from five different views spanned between the frontal and profile views. Moreover, a simple VSR system has been developed and tested on the database to provide some baseline performance.

Head Pose and Eye Gaze Estimation

As large datasets with well-defined gaze directions are desired for the researches in eye gaze estimation, we collected the Oulu Multi-pose Eye Gaze Dataset in order to facilitate related researches. The Oulu Multi-pose Eye Gaze Dataset finally includes 200 image sequences from 50 subjects (For each subject it includes four image sequences). Each sequence consists of 225 frames captured when people are fixating on 10 targeting points on the screen. The first three sequences of each subject are captured under three fixed head poses, namely 0 (the frontal) and ±30 degree respectively. The last sequence is in a free pose style. It forms a strong basis for considering the eye gaze estimation in a finer manner; say continuous manner, rather than just several discrete angles as previous researches do. Moreover, we provided baseline results on our dataset by evaluating the popular approaches on eye gaze estimation. To discuss the influence of different head poses in gaze estimation, we further designed a set of experiments, which randomly pick fifty percent of subjects in one pose as the training set and the left subjects in other poses as the testing set. It shows that the performance under the same pose is much higher than the one under different poses. It means the eye gazes are distinctly different when head poses vary. It confirms our estimation that multi-pose gaze estimation is highly challenging. Fortunately, our dataset provides an opportunity of in-depth investigation in this issue.

Sample of the normalized and cropped eye gaze images in a resolution of 30 × 150 pixels from three fixed head poses respectively.
As a promising technology, visual attention analysis contributes to various computer vision based applications, such as object detection and image segmentation. Since eye movement reveals the regions of interest (ROI) of the human visual system, it is widely utilized in researches regarding visual attention understanding. Hence, to facilitate the research in visual attention analysis, we design and establish a new task-driven eye tracking dataset of 47 subjects. Inspired by psychological findings that human visual behavior is tightly dependent on the executed tasks, we carefully design specific tasks in accordance with the contents of 111 images covering various semantic categories, such as text, facial expression, texture, pose, and gaze. It results in a dataset of 111 fixation density maps and over 5,000 scanpaths. Moreover, we provide baseline results of thirteen state-of-the-art saliency models. Furthermore, we hold discussions on important clues on how tasks and image contents influence human visual behavior. This task-driven eye tracking dataset with the fixation density maps and scanpaths will be made publicly available.

(a) Examples of watching materials and the created fixation density maps. (b) Example of a scanpath. The circles are fixation positions and the length of radius correlates to the duration between fixations. The line shows the direction of fixation shifts and the circle without cross symbol is the first fixation point. The first 5 fixations are marked in this image.

Affective Human-Computer Interaction

A paper describing our Minotaurus system developed for affective human-robot interaction in smart environments was published in 2014 in Cognitive Computation journal.

Vision Systems Engineering

Vision systems engineering research aims to identify attractive computing approaches, architectures, and algorithms for industrial machine vision systems. In this research, solutions ranging from low-level image processing even to equipment installation and operating procedures are considered simultaneously. The roots of this expertise are in our visual inspection studies in which we met extreme computational requirements already in the early 1980’s, and we have contributed to the designs of several industrial solutions. We have also applied our expertise to applications intended for embedded mobile platforms.

In 2014, we created a demo system that showcases different face analysis methods including face recognition, gender recognition, facial expression recognition and heart rate measurement. In 2015-early 2016, key parts of this system are ported to work on Vuzic M100 smart glasses. Software uses only local processing and does not rely on network connectivity to offload computation to servers. Face analysis technology on wearable devices enables many different kind of services. Familiar persons could be associated with extra information that is displayed when they are detected. In the future, software could also automatically analyze surroundings, e.g. for the visually impaired and use other modalities to convey this information for the user.

Face analysis with Vuzic M100 smart glasses.

Biomedical Image Analysis

In recent years, increasing resolving power and automation of biomedical imaging systems have resulted in an exponential growth of the image data. Manual analysis of these data sets is extremely labor intensive and hampers the objectivity and reproducibility of results. Hence, there is a growing need for automatic image processing and analysis methods. In CMV, our aim has been to apply modern computer vision techniques to biomedical image analysis which is one of our emerging research areas.

Automated cell classification in Indirect Immunofluorescence (IIF) images has potential to be an important
tool in clinical practice and research. Recently, classification of Human Epithelial Type 2 (HEp-2) cell images has attracted great attention. However, the HEp-2 cell classification task is quite challenging due to large intra-class and small inter-class variations. We proposed an effective approach for the automatic HEp-2 cell classification by combining multi-resolution co-occurrence texture and large regional shape information. In our approach, we: a) capture multi-resolution co-occurrence texture information by a novel Pairwise Rotation Invariant Co-occurrence of Local Gabor Binary Pattern (PRICoLGBP) descriptor, b) depict large regional shape information by using an Improved Fisher Vector (IFV) model with RootSIFT features which are sampled from large image patches in multiple scales, and c) combine both features. We evaluated systematically the proposed approach on the IEEE International Conference on Pattern Recognition (ICPR) 2012, the IEEE International Conference on Image Processing (ICIP) 2013 and the ICPR 2014 contest data sets. The proposed method based on the combination of the introduced two features outperforms the winners of the ICPR 2012 contest using the same experimental protocol. Our method also greatly improves the winner of the ICIP 2013 contest under four different experimental setups. Using the leave-one-specimen-out evaluation strategy, our method achieves comparable performance with the winner of the ICPR 2014 contest that combined four features. This work is published in IEEE Journal of Biomedical and Health Informatics.

In another paper, published in Pattern Recognition Letters, we analyzed the importance of the pre-processing, and more specifically the role of Gaussian Scale Space (GSS) theory as a pre-processing approach for the HEp-2 cell classification task. We validated the GSS pre-processing under the Local Binary Pattern (LBP) and the Bag-of-Words (BoW) frameworks. Under the BoW framework, the introduced pre-processing approach, using only one Local Orientation Adaptive Descriptor (LOAD), achieved superior performance on the Executable Thematic on Pattern Recognition Techniques for Indirect Immunofluorescence (ET-PRT-IHF) image analysis. Our system, using only one feature, outperformed the winner of the ICPR 2014 contest that combined four types of features. Meanwhile, the proposed pre-processing method is not restricted to this work; it can be generalized to many existing works.

In our recent work, we also presented a framework for classification of HEp-2 cell images using convolutional neural networks (CNNs). Previous state-of-the-art methods show classification accuracy of 75.6% on a benchmark dataset. We conducted an exploration of different strategies for enhancing, augmenting and processing training data in a CNN framework for image classification. We demonstrate how training data affects classification accuracy of cell classification. We found that additional real data-augmentation is incredibly helpful and domain specific pre-training still main-

In last few years, advances in microscopy techniques have enabled the investigation of dynamic processes at increasing temporal and spatial resolution. This has produced large quantities of imaging data, which cannot be fully analyzed manually. This has increased the importance of automatic analysis methods, most of which depend heavily on accurate cell segmentation and tracking.

Microscopic images can be very challenging when cell density is high due to frequent interaction of cells with each other. Often there is not enough information in a single frame to make the correct decision about segmentation and tracking. In these situations it helps to consider content of adjacent frames when making decisions, which can be computationally very expensive for long dense sequences. We have developed a greedy joint cell segmentation and tracking method which overcomes this challenge.

Our method uses multiple filter banks to detect cells and uses watershed to split cell clusters and obtain cell proposals. It then creates a hierarchical forest from these cell proposals. The figure below shows two trees in the forest (a). Cells in microscopic sequences can go through few events, the probabilities of which are represented by nodes in a super-node (b). Proposal super-nodes in adjacent frames are connected with each other to create a directed acyclic graph (c). Tracks within this graph are found by iteratively finding the shortest path, which provides cell segmentations and tracks.
A joint cell segmentation and tracking method using cell proposals.

Exploitation of Results

Many researchers have adopted and further developed our methodologies. Our research results are used in a wide variety of different applications around the world. For example, the Local Binary Pattern methodology and its variants are used in numerous image analysis tasks and applications, such as biomedical image analysis, biometrics, industrial inspection, remote sensing and video analysis. The researchers in CMV have actively published the source codes of their algorithms for the research community, and this has increased the exploitation of the results.

The results have also been utilized in our own projects. For example, we have collaborated with Prof. Tapio Seppänen’s Biomedical Engineering Group in the area of multimodal emotion recognition for affective computing, combining vision with physiological signals. Together with Prof. Osmo Tervonen from Oulu University Hospital we have been carried out research on classifying thorax images using computer vision and deep learning methods.

Most of our funding for both basic and applied research comes from public sources such as the Academy of Finland and Tekes, but besides these sources, CMV also conducts research by contract which is funded by companies. In this way, our expertise is being utilized by industry for commercial purposes, and even in consumer products, like mobile devices.

The CMV has actively encouraged and supported the birth of research group spin-outs. This gives an opportunity for young researchers to start their own teams and groups. Side results are the spin-out enterprises. According to our experience, their roots are especially in the strands of “free academic research”.

Future Goals

Our results from 2015 are very positive, for example the number of publications in major forums has clearly increased. Having two new FiDiPro projects for distinguished scientists from abroad will make an exciting progress possible also in coming years. We will continue to sharpen our strategies to meet the future demands and ensure enough research funding in an increasingly tough competition. We plan to carry out well focused cutting-edge research, for example, on novel image and video descriptors, multimodal face analysis and biometrics, multimodal analysis of emotions, 3D computer vision, biomedical image analysis, and energy-efficient architectures for embedded vision systems. We also have plans to further deepen our collaboration with international and domestic partners. We plan to participate in new European project proposals, and continue applying funding for breakthrough research from the Academy of Finland and the European Research Council (ERC). Close interaction between basic and applied research has always been a major strength of our research unit. The scientific output of the CMV has been increasing significantly in recent years. With this we expect to have much new potential for producing novel innovations and exploitation of research results in collaboration with companies and other partners.

Personnel

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>professors</td>
<td>5</td>
</tr>
<tr>
<td>senior research fellows</td>
<td>4</td>
</tr>
<tr>
<td>postdoctoral researchers</td>
<td>11</td>
</tr>
<tr>
<td>doctoral students</td>
<td>16</td>
</tr>
<tr>
<td>other research staff</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>42</td>
</tr>
<tr>
<td>person years for research</td>
<td>34</td>
</tr>
</tbody>
</table>

External Funding

<table>
<thead>
<tr>
<th>Source</th>
<th>EUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>1 401 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>142 000</td>
</tr>
<tr>
<td>total</td>
<td>1 543 000</td>
</tr>
</tbody>
</table>

Doctoral Theses


Selected Publications


