Issue N° 3 / October 2006

Editor : Céline Aymon, celine.aymon@idiap.ch

DIRAC

Detection and Identification of Rare Audiovisual Cues

Newsletter

www.diracproject.org

DIRAC c/o IDIAP Research Institute, Simplon 4, P.O. Box 592, CH-1920 Martigny
info@diracproject.org - www.diracproject.org

Contents

News

Join the MAIA Workshop

Challenging Brain Computer Interfaces: November, 9-10, 2006
http://www.maia-project.org/workshop-2006.php
More information on page 4.

Cover Story

DIRAC First Application Scenario

It is a common human experience that a number of actions in our lives are carried out almost automatically and without requiring extensive mental efforts. Imagine a daily routine of going to buy your morning newspaper across the street. You put your shoes and your hat on, open the door of your flat, take a lift down with the usual boring background music, get out, greet the lady in the shop in front of your house and hear her reply, avoid the old car parked on the sidewalk ever since you moved into your house, cross the street when the green light comes on, pay 3.80 CHF asking price for it, and walk the same route back Not much of your mental energy spent so far, you are ready to start reading about what is new in the world.

Now, imagine that one day one of the items along your way has changed. Perhaps the shop owner asks 4.20 CHF for the newspaper. Your cognitive system signals alarm! You examine the salesman's voice (sounds the same as every day, go to the next check), look at him (not a robber who just robbed the shop but the same man you see every day, this danger is eliminated), look at the newspaper (the same newspaper - nothing wrong, this danger eliminated), ask the salesman and hear his reply that the price just went up. Your inner mental model of the world gets updated; the next day you hear 4.20 CHF asking price and you pay it without any alarm.

Research efforts in the DIRAC consortium are aimed at identifying and describing of such unexpected events. Their inherent low prior probability, and their low occurrence or even complete absence in training data that are used in the design of the machine, present a set of new and interesting research challenges. One of the important goals of the DIRAC consortium in its first year is to arrive at exemplar application scenarios that would serve as the focal point of our research. After extensive discussions among partners, one particular application scenario is starting to emerge. It calls for the design of an audio-visual device that would be able to learn daily routines of its user and issue alarms when the routine is violated, could help to identify dangerous deviations from user's behaviour in the care for the elderly, or be of utility to law enforcement personnel in their daily routine patrol. When stationary, the device can be used e.g. for monitoring of audio-visual environments in the care for the elderly or surveillance applications, or it may attempt to identify and describe information-rich unexpected elements in human group communication and behaviour. Mounted on wheel-chair, it can help elderly in getting around their daily tasks. Ultimately, we are aiming for a wearable device on the user's body where we will need to address a number of additional problems arising from the movements of optical and acoustic sensors in realistic noisy environments, among them the need for fast audio-visual identification of changes in the scene (a problem that is attracting the attention of the hearing aid industry), the need for accurate identification and description of unexpected words in speech (that is of interest in machine surveillance), the need for detailed description of 3-D visual scenes captured by moving visual sensors (that is required in a number of practical industrial applications), the need for learning and categorization from small samples, the need for continuous model adaptation, and the need for efficient fusion of information from various information sub-streams (problems which are of a great research interest to machine learning community).

Detection and identification of Rare Audio visual Cues (DIRAC) is an Integrated project funded by the EC's 6th Framework Program, managed by IDIAP (CH).

Issue N° 3 / October 2006

Editor : Céline Aymon, celine.aymon@idiap.ch

Detection and Identification of Rare Audiovisual Cues

Newsletter

The Center for Machine Perception, Czech Technical University, Prague

The Center for Machine Perception (CMP, cmp.felk.cvut.cz) at the Czech Technical University, Prague is a research unit focusing on computer vision, pattern recognition, and mathematics. The Center for Machine Perception comprises 25 professors and researchers and about 20 PhD students at the Department of Cybernetics (EC Center of Excellence) of the Czech Technical University Prague. It participates in a number of EU-funded and national projects.

Research interests of the Center for Machine Perception span from basic research on algebra, quantum and fuzzy logic, digital topology, statistical estimation, geometry of image formation, omnidirectional vision, non-classical cameras, object recognition in images, and 3D reconstruction from images, to applied research and development of object tracking for Unmanned Aerial Vehicles, license plate recognition for traffic monitoring and car speed measurement, and omnidirectional visual navigation for tucks and trailers.

CMP Team : from left to right
Akihiko Torii, Hynek Bakstein, Tomas Pajdla, Zuzana Kukelova, Michal Havlena

Tomas Pajdla's group focuses on understanding, modeling, and design of geometry of image formation and 3D scenes understanding from images. In past we concentrated mainly on geometrical aspects of computer vision, visual robot control, eye-hand calibration and coordination, precise digital optical measurements, photogrammetry, and robot navigation using vision. Our main interest and results were in panoramic and non-classical imaging. We introduced epipolar geometry of panoramic cameras, investigated the use of panoramic images for robot localization, contributed to studies of panoramic mosaics, and proposed to study omni-directional cameras that do not have a center of projection but have a generalization of the epipolar geometry. Recently, we contributed to image matching and solving wide-baseline stereo correspondence problem. We are also becoming more interested in combinations of algebraic and statistical techniques aiming at identifying models and events in very cluttered data.

In DIRAC, CMP leads WP-1, developing vision and acoustic sensors and task-driven optimized signal processing. CMP also concentrate on applications of optimized vision sensors and for research and development of advanced signal processing techniques suited for extracting information from images. An examples is image features detection for matching and visual object and scene recognition. Other major contributions of CMP are in WP-3 where we develop a framework for 3D reconstruction from omni-directional images using «cognitive feedback» and in WP-6 where we concentrate on integrating omnidirectional vision into the DIRAC demonstrator and on demonstrating omnidirectional camera tracking.

Omnidirectional imaging is very useful in visual surveillance, visual events detection and recognition, autonomous vehicle navigation, and scene modeling from images. A particularly important problem related to omnidirectional imaging is modeling of image projection and estimation of its parameters. Hynek Bakstein develops non-classical imaging sensors, their projection models, and methods for processing omnidirectional images.

Matching of instant images to visual representation of the scene and camera motion tracking can be used to detect unexpected changes in the scene appearance and sudden changes of motion pattern of a pedestrian walking in a city. No matching based purely on image similarities is perfect. Fortunately, models and constraints can be fit to even very cluttered data by pseudo-random generation and verification of hypotheses. Michal Havlena develops camera motion tracking and 3D structure estimation from omnidirectional images. He also looks at improvements of the pseudo-random fitting to incorporate additional constraints on matching efficiently.

Estimating models from image data often calls for solving systems of algebraic equations. Recently, algorithms for solving general algebraic systems have been developed but they may have even double exponential computational complexity is applied blindly. To solve practical problems, special algorithms implementing «shortcuts» in solutions need to be developed. Zuzana Kukelova is developing specialized solvers for the systems arising in omnidirectional camera autocalibration, camera motion estimation, and detection and description of geometrical image features.

High level knowledge can be used to drive low level feature detection and hypotheses generation to improve robustness and efficiency of processing through a «cognitive feedback». Akihiko Torii works on image matching driven by higher level knowledge. Our current research develops along two lines. First, we try to learn which feature detectors succeed in which parts of images to quickly rule out those which have low chance of success. Secondly, we investigate how to influence hypotheses generation by already extracted knowledge to generate promising hypotheses early.

DIRAC
Detection and Identification of Rare Audiovisual Cues

Issue N° 3 / October 2006

Editor : Céline Aymon, celine.aymon@idiap.ch

Newsletter

Non-traditional audio-visual sensors in DIRAC

OMNIDIRECTIONAL IMAGING AND HEARING AID MICROPHONE ARRAYS

DIRAC's focus on audio-visual integration and human-inspired cognition requires adequate sensors and sensory processing, an area to which two groups in the consortium contribute their special expertise. Panoramic cameras for acquiring visual input are a specialty of Tomas Pajdla's group at the Czech Technical University in Prague, Czech Republic, and human-centered audio acquisition and processing are a core competence of Birger Kollmeier's Medical Physics group at the Carl von Ossietzky-University in Oldenburg, Germany.

Body mounted hearing aid microphones

The audio sensors DIRAC uses are «dummy hearing aids», i.e., microphone arrays worn behind the ears in the casing of a hearing aid, but without any built-in signal processing. We are presently using a six-channel array with three microphones placed behind each ear. Fig. 1 shows a close-up of a single three-channel casing in situ. The choice of this device has some advantages specific to DIRAC's scientific and application goals: With sensors at both sides of the head, part of the array geometry resembles that of human hearing which heavily draws on the shading effect of the head and on the relatively large distance between the ears. At the same time, human-inspired processing can be combined with more traditional techniques (e.g., beam-forming) that are taylored towards the much smaller inter-microphone distances found within each microphone triplet. A thin wire connects the hearing-aid dummies to a firewire sound card and a laptop computer. The whole system is battery powered and highly portable which has been used in recordings made with a person walking around in a city centre and riding a bicycle while recording sound. Hence, these sensors ideally fit DIRAC's envisaged application of a cognitive audio-visual aid.

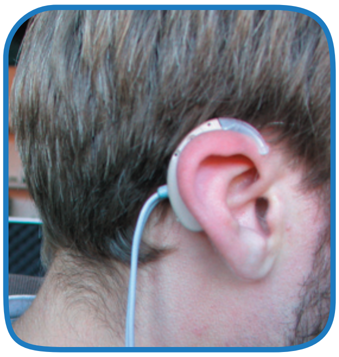

Figure 1

Present recordings are performed with the goal of recording «benchmark» data for testing of our algorithms under different controlled environmental conditions. Therefore, we systematically alter a number of acoustic parameters. E.g., degree of reverberation is varied between the prototypical «free field» situation, recorded in the anechoic chamber of Oldenburg University, and the increased reverberation of a typical office room and larger spaces such as conference room or train station. Both indoors and outdoors scenes with a varying number of active sound sources are taken into account. Another important parameter is the dynamics of the acoustic environment: Is it stationary with fixed listener and fixed sound sources (typical in an office environment), or is it characterized by moving sound sources and possibly a moving listener (such as a person standing or walking in the street)?

For some of these situations it is possible to record sound sources and transfer functions separately and later mix them digitally to obtain a «virtual» acoustic scene. This gives us very controlled access to the acoustic parameters our algorithms have to deal with and thereby facilitates a detailed analysis of their performance, while still working on fully realistic data.

Panoramic camera sensors

Omnidirectional vision provides peripheral vision which is essential for reacting to fast, unpredictable, and unusual visual events. The main issues of every image acquisition are the view-angle, resolution, frame rate, image quality and size. For a particular task, building an optimized device is still a matter of special development. It is one of the goals of DIRAC Workpackage 1 to design and realize an omnidirectional acquisition sensor suitable for DIRAC demonstrators. A first high quality though rather «heavy-weight» prototype has been built based on a pair of externally fired Kyocera M410R cameras with Nikon FC-E9 183 degrees field of view lenses capable of acquiring 4 megapixel images at three frames per second, shown in Fig. 2.

DIRAC's AWEAR demonstrator, however, will have to rely on much smaller and cheaper devices. Therefore we are working on putting together a more «light-weight» device with acceptable parameters. After a number of experiments with various optical components we have found a miniature light-weight fish-eye lens, Sunex DSL-215 fisheye lens, with 185 degrees field of view and dimensions that fit into the 2 x 2 x 2 cm cube. This device has lower resolution, higher frame-rate but only about 130 degrees rectangular field of view. The next step in our development is to integrate the lens with a suitable imager to get full view angle.

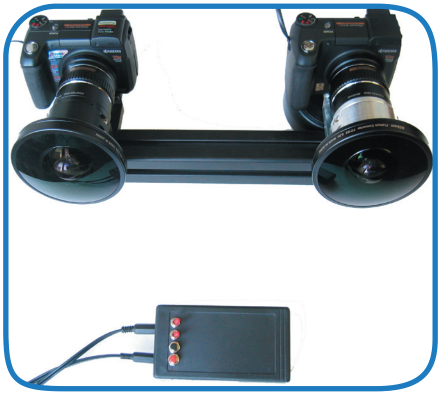

Figure 2

The price to pay for large view angles is unusual image projection and therefore a need to develop non-standard image processing and understanding techniques. Camera calibration allowing to generate perspective cutouts from circular omnidirectional images is a particularly important problem related to omnidirectional imaging. Our methods allow to perform calibration either by using images of a known calibration model or by using correspondences extracted by image matching based on salient feature recognition. Developing auto-calibration and image matching techniques for non-perspective imaging is one topic of DIRAC Workpackages 1 and 3. The main issues here are related to feature detection, description and recognition in omnidirectional images and efficient techniques for estimating correct models from data contaminated by gross errors. In such a case, encountering a correctly estimated model is a very rate event.

First integrated audio-visual data acquisition

The audio and visual recording setups have been combined at the recent DIRAC meeting in Leuven and first audio-visual data has been recorded indoors and outdoors on this occasion. These recordings are providing us with valuable initial experience for audio-visual data recording and processing. The construction of a fully integrated audio-visual recording device is the next logical step which is already under way.

News & Events

Challenging Brain Computer Interfaces: Neural Engineering Meets Clinical Needs in Neurorehabilitation

Rome Italy, November, 9-10, 2006

Last years have witnessed advances in Brain-Computer Interfaces (BCI), but how far is this new field from clinical practice? The goal of the workshop is to draw the current and future scenarios involving the application of advanced neural engineering techniques to interpret brain signals for clinical use in the rehabilitation context.

The presentations will consist of a series of invited talks and poster presentations. Some of the major groups in BCI pursuing clinical applications of this technology will report their experience. Also, the view of clinicians involved in neurorehabilitation programs will complete the picture. Finally, the European MAIA project will report their achievements in non-invasive brain-controlled robots. Altogether the workshop will address how ultimate neural engineering techniques could meet the challenge of neurorehabilitation.

For more information please visit
www.maia-project.org/workshop-2006.php

DIRAC's Publications

(http://www.diracproject.org/publications/)

Speaker Localization for Microphone Array-Based ASR: The Effects of Accuracy on Overlapping Speech

Hari Krishna Maganti and Daniel Gatica-Perez

Eighth International Conference on Multimodal Interfaces (ICMI'06), November 2-4, 2006, Banff, Canada

Accurate speaker location is essential for optimal performance of distant speech acquisition systems using microphone array techniques. However, to the best of our knowledge, no comprehensive studies on the degradation of automatic speech recognition (ASR) as a function of speaker location accuracy in a multi-party scenario exist. In this paper, we describe a framework for evaluation of the effects of speaker location errors on a microphone array-based ASR system, in the context of meetings in multi-sensor rooms comprising multiple cameras and microphones. Speakers are manually annotated in videos in different camera views, and triangulation is used to determine an accurate speaker location. Errors in the speaker location are then induced in a systematic manner to observe their influence on speech recognition performance. The system is evaluated on real overlapping speech data collected with simultaneous speakers in a meeting room. The results are compared with those obtained from close-talking headset microphones, lapel microphones, and speaker location based on audio-only and audio-visual information approaches.

Learning a Kernel Function for Classification with Small Training Samples

Tomer Hertz, Aharon Bar-Hillel, and Daphna Weinshall

23rd International Conference on Machine Learning (ICML), June 25-29 2006, Pittsburgh Pennsylvania, USA

When given a small sample, we show that classification with SVM can be considerably enhanced by using a kernel function learned from the training data prior to discrimination. This kernel is also shown to enhance retrieval based on data similarity. Specifically, we describe KernelBoost - a boosting algorithm which computes a kernel function as a combination of 'weak' space partitions. The kernel learning method naturally incorporates domain knowledge in the form of unlabeled data (i.e. in a semi-supervised or transductive settings), and also in the form of labeled samples from relevant related problems (i.e. in a learning-to-learn scenario). The latter goal is accomplished by learning a single kernel function for all classes. We show comparative evaluations of our method on datasets from the UCI repository. We demonstrate performance enhancement on two challenging tasks: digit classification with kernel SVM, and facial image retrieval based on image similarity as measured by the learnt kernel.

Doron Feldman and Daphna Weinshall

Motion Segmentation Using an Occlusion Detector

Workshop on Dynamical Vision, in 9th European Conference on Computer Vision (ECCV), May 7-13 2006, Graz Austria

We present a novel method for the detection of motion boundaries in a video sequence based on differential properties of the spatio-temporal domain. Regarding the video sequence as a 3D spatio-temporal function, we consider the second moment matrix of its gradients (averaged over a local window), and show that the eigenvalues of this matrix can be used to detect occlusions and motion discontinuities. Since these cannot always be determined locally (due to false corners and the aperture problem), a scale-space approach is used for extracting the location of motion boundaries. A closed contour is then constructed from the most salient boundary fragments, to provide the final segmentation. The method is shown to give good results on pairs of real images taken in general motion. We use synthetic data to show its robustness to high levels of noise and illumination changes; we also include cases where no intensity edge exists at the location of the motion boundary, or when no parametric motion model can describe the data.

Integrating Recognition and Reconstruction for Cognitive Traffic Scene

Bastian Leibe, Nico Cornelis, Kurt Cornelis and Luc Van Gool

28th Annual Symposium of the German Association for Pattern Recognition (DAGM), September 12-14, 2006, Berlin Germany

This paper presents a practical system for vision-based traffic scene analysis from a moving vehicle based on a cognitive feedback loop which integrates real-time geometry estimation with appearance-based object detection. We demonstrate how those two components can benefit from each other's continuous input and how the transferred knowledge can be used to improve scene analysis. Thus, scene interpretation is not left as a matter of logical reasoning, but is instead addressed by the repeated interaction and consistency checks between different levels and modes of visual processing. As our results show, the proposed tight integration significantly increases recognition performance, as well as overall system robustness. In addition, it enables the construction of novel capabilities such as the accurate 3D estimation of object locations and orientations and their temporal integration in a world coordinate frame. The system is evaluated on a challenging real-world car detection task in an urban scenario.

Machine recognition of speech consistent with some properties of auditory cortical receptive fields

Hynek Hermansky

International Conference on the Auditory Cortex" The Listening Brain", September 17-21, 2006 Nottingham, United Kingdom

Automatic recognition of speech consistent with some properties of auditory cortical receptive fields Features describing instantaneous shape of an auditory-like modified short-term spectrum of speech together with features describing temporal dynamics of spectral envelopes form the basis of most conventional techniques in automatic speech recognition (ASR). We demonstrate advantage of abandoning the spectral envelope altogether, replacing it by frequency-specific posterior probabilities of speech-related events, derived by projecting the speech time-frequency plane on 2-D (time-frequency) basis functions that are consistent with some properties of auditory cortical receptive fields. That then leads to a new data-driven feature extraction technique which derives ASR features related to posterior probabilities of context-independent phoneme classes. These features yield advantage in phoneme recognizer that in turn can be applied for detection and identification of low-prior-probability words.