IM2 Summer Institute Schedule (14.11.05-17.11.05)

Monday	Tuesday	Wednesday	Thursday
9:00-9:30 Welcome to Summer Institute	**8:30-10:30**: Workshop part I Speaker: Veenhuys David, "Marketing Yourself", A seminar about how to apply modern marketing to job hunting and career development in the research area.	**9.30-12.30** Workshop Part II Speaker: Veenhuys David, "Marketing Yourself", A seminar about how to apply modern marketing to job hunting and career development in the research area.	**9.30-12.30** IM2 Phase II, Presentation of new IPs and management structure: 9:30-9:45 Introduction & Presentation of Phase II management structure, Herve Bourlard/Jean-Albert Ferrez 9:45-10:00 IM2.DMA: Database management and meeting analysis, Andrei Popescu-Belis 10:00-10:15 IM2.AP: Audio Processing, John Dines 10:15-10:30IM2.VP: Visual/Video Processing, Jean-Philippe Thiran
9:30-10:00 Jonas Richiardi and Krzysztof Kryszczuk, "Estimating reliability in uni and multimodal biometric user verification", IM2.ACP			
10:00-10:30 Guillaume Lathoud, " Unsupervised Speaker Detection-Localization", IM2.SP			
10:30-11:00 Coffee Break	10:30-10:45 Coffee Break	10:30-10:45 Coffee Break	10:30-10:45 Coffee Break
11:00-11:30 Serhiy Kosinov, "Large margin multiple hyperplane classification for content-based multimedia retrieval", IM2.IIR	10:30-12:30: Workshop part I Speaker: Veenhuys David, "Marketing Yourself", A seminar about how to apply modern marketing to job hunting and career development in the research area.	10.30-12.30 Workshop Part II Speaker: Veenhuys David, "Marketing Yourself", A seminar about how to apply modern marketing to job hunting and career development in the research area.	10:45-11:00 IM2.MPR: Multimodal Processing and recognition, Samy Bengio 11:00-11:15 IM2.MCA: Multimodal context abstraction, Stefan Marchand-Maillet 11:15- 11:30 IM2.HMI: Human-machine integration, Pierre Wellner 11:30-11:45 IM2.ISD:Integration software and research demonstration, Mike Flynn 11:45-12:00 IM2.BMI: Brain machine interaction, José del R. Millán 12:00-12:30 Closing of Summer Institute 2005, Herve Bourlard
11:30-12:00 Hamed Ketabdar, "Using More Informative Posterior Probabilities for Speech Recognition", IM2.SP			
12:00-12:30 Anna Buttfield, "Online Learning for the IDIAP Brain-Computer Interface", IM2.MI			
12:30-13:30 Lunch	**12:30-13:30 Lunch**	**12:30-13:30 Lunch**	
13:30-14:30 **Invited Speaker**: Gerhard Rigoll, "Multimodal Interaction in Smart Environments".	13:30-14:00 Dalila Mekhaldi, "Thematic alignment of static Documents with Meeting dialogs", IM2.DI	13:30- 14:00 **Guest Speaker TBA** or Pierre W. Ferrez, "Automatic Detection of Interaction Errors from EEG" IM2.MI	

14:30-15:30 Set up poster	14:00-14:30 Dong Zhang, "Learning influence among human interactions", IM2.MI	14:00-14:30 Florent Monay/Pedro Quelhas, "Constructing visual models with local descripors and latent aspects", IM2. IIR
15:30-18:00 **Poster Session with NCCR Review Panel participating and Coffee Break**	14:30-15:00 Tobias Kaufmann, "Using Rule-based Knowledge to Improve LVCSR", IM2. SP	14:30-15:00 Gianluca Monaci, "Analysis of Multimodal Signals Using Redundant Representations", IM2. SA
	15:00-15:30 Nicolas Moenne-Loccoz, "Interactive Retrieval of Video Sequences from Local Feature Dynamics", IM2. IIR	15:00-15:30 Ardhendu Behera, "DocMIR, an Automatic Document-based Indexing System for Meeting Retrieval", IM2.DI
	15:30-16:00 Coffee Break	15:30-18:00 **Poster Session with IM2 Scientific and Industrial Advisory Board participating & Coffee Break**
	16:00-16:30 Mark Barnard, " Event recognition in sports videos using layered HMMs", IM2. MI	
	16:30-17:00 Mael Guillemot, "A Meeting Recording Corpus", IM2. IP	
	17:00-17:30 Siley Ba, "Probabilistic Models for Head Pose Tracking in Meetings" IM2.SA	
	17:30-18:00 Marita Ailomaa, "Accessing Multimedia Meeting Data Multimodally", IM2.MDM	
19:00 Dinner	**19:00 Dinner**	**19:00 Dinner**

Abstract of IM2 Summer Institute Talks

Monday

Jonas Richiardi and Krzysztof Kryszczuk, “Estimating reliability in uni and multimodal biometric user verification”,

Many areas of pattern recognition make use of measures that express the uncertainty in a classifier's output, at the measurement (score) or decision levels. Speech recognition systems often output confidence levels with recognized words, dialogue systems use them to govern actions and repair strategies, and handwriting recognition applications make use of estimates of the segmentation module's output to drive feedback mechanisms. In biometric authentication, the idea that the uncertainty in the output of classifiers should be quantified has gained ground in the last few years.

One of the most important factors affecting the performance of biometric verification systems (apart from lack of training data and improper acquisition) is the mismatch between training and testing conditions. When noise is present, the recognition performance can drop dramatically.

This talk will look at measures of biometric verification decision reliability, and demonstrate some ways to react to low reliabilities in biometric identity verification. We define decision reliability as ``the posterior probability of taking a correct decision given available evidence", where evidence can come from the classifier domain (scores, posteriors, n-best lists), feature domain, or signal domain (e.g. signal-to-noise-ratio for speech or based on pixel intensity variance for face). We present our probabilistic framework for eliciting reliability measures, which combines several sources of evidence in a scalable fashion. We also demonstrate usage of the measure in a bimodal setting (speech and face). IM2.ACP

Guillaume Lathoud, “Unsupervised Speaker Detection-Localization”,

Spontaneous multi-party conversations have highly dynamic and changing characteristics. "Dynamic" because speaker turns are usually very short, and speakers often interrupt each other. "Changing" because people may move or even physically leave during a conversation. In this context, speech processing applications are required to (1) detect accurately each spoken word, (2) adapt to various unknown conditions automatically, for which prior knowledge is limited or unavailable. In this talk, we'll propose a generic unsupervised detection approach. The advantage of this approach over traditional training/testing approaches is shown on a microphone array detection-localization task, including results on real meeting room recordings. IM2.SP

Serhiy Kosinov, “Large margin multiple hyperplane classification for content-based multimedia retrieval”,

In this note, we consider the asymmetric classification problem setting, often encountered in content-based multimedia retrieval performed as a ``one-against-all" classification scheme. The essence of the proposed technique is to increase the number of hyperplanes used in an optimal separating hyperplane classifier, so as to favor the under-represented class. Such a distinction that singles out a certain target class from the rest of the data, when modeled explicitly, has been previously shown to improve classification accuracy for undersampled and unbalanced data sets. While being applicable in the general classification scenario, the proposed method is designed to further exploit the asymmetry of the classification problem at hand. A preliminary empirical evaluation of the proposed method provides encouraging results, which warrants further investigation. IM2 .IIR

Hamed Ketabdar, “Using More Informative Posterior Probabilities for Speech Recognition”,

In this talk, we present initial investigations towards boosting posterior probability based speech recognition systems by estimating more informative posteriors taking into account acoustic context (e.g., the whole utterance), as well as possible prior information (such as phonetic and lexical knowledge). These posteriors are estimated based on HMM state posterior probability definition (typically used in standard HMMs training). This approach provides a new, principled, theoretical framework for hierarchical estimation/use of more informative posteriors integrating appropriate context and prior knowledge. In the present work, we used the resulting posteriors as local scores for decoding. On the OGI numbers database, this resulted in significant performance improvement, compared to using MLP estimated posteriors for decoding (hybrid HMM/ANN approach) for clean and more specially for noisy speech. The system is also shown to be much less sensitive to tuning factors (such as phone deletion penalty, language model scaling) compared to the standard HMM/ANN and HMM/GMM systems, thus practically it does not need to be tuned to achieve the best possible performance. IM2.SP

Anna Buttfield, “Online Learning for the IDIAP Brain-Computer Interface”, IM2.MI

The goal of a brain computer interface (BCI) is to enable a user to give commands to a computer simply by thinking, through translating neural activity into command signals. One method of achieving this non-invasively is to measure the electrical activity of the brain (EEG) while the subject is performing a set of tasks ? for example, imagining left and right hand movement, and a language task. This is challenging for a number of reasons, including the noisiness of the EEG signals, and

the concentration required from the subject. In addition the EEG signals change over time, both within a single session and between sessions, due to factors such as user fatigue or change in strategy, and variations in signal noise. This means that while we can train a classifier on data from previous sessions, it will probably not be optimal for subsequent sessions. It is also desirable that the classifier has the ability of adapt itself in order to track changes in the signal throughout a session. This is particularly important in the initial training period, where we would like to allow the classifier to adapt rapidly and give the subject appropriate feedback, allowing him or her to select the mental strategy that works best. We are investigating methods of incorporating online learning into the IDIAP BCI, which takes the EEG signal from a 32 or 64 electrode cap and classifies the signal with a statistical Gaussian classifier. This system has been proven effective at differentiating between two or three mental states at a time (for example, imagination of left and right hand movement and a language-based task). This has been applied to problems such as operating simple computer games, and navigating a miniature robot around a model indoor environment with rooms and corridors. Work so far has involved the specific problem of the training phase, where the subject is told which command to give and so the real target is always known, making it a supervised learning task. We have been investigating various gradient-descent based learning algorithms that can be applied to the Gaussian classifier in an on-line manner. One such method is stochastic meta-descent (Schraudolph, 1999), which accelerates training by using local learning rate adaptation. IM2.MI

Invited Speaker:

Gerhard Rigoll, “Multimodal Interaction in Smart Environments”.

In this talk, multimodal human-machine interaction will be presented from the general viewpoint of "smart environments". Smart environments have the common feature that they are equipped with a number of sensors that enable the environment to communicate with the humans that are in this environment. Thus, advanced human-machine interaction using more than one single modality should be an essential capability of smart environments. The talk will present some overview of the activites in multimodal HCI as pursued at Munich University of Technology, with some special emphasis on selected topics, such as speech- and vision-based emotion recognition, face profile recognition and new approaches for person tracking. Some of the most popular smart environments are probably the smart home and the smart meeting room scenario. In contrast to that, this talk will highlight some activities in alternative smart environments, namely in the area of "smart cockpits". Some concrete results will be presented from actual projects with prominent automobile manufacturers in the area of smart car indoor environments and from a European project in the aeronautics sector dealing with smart air cabins for the new generation of the Airbus.

Gehrard Rigoll is Professor at the Institute for Human-Machine Communication Munich University of Technology, Germany and Member of the IM2 Review Panel

Tuesday Afternoon

Dalila Mekhaldi, "Thematic alignment of static Documents with Meeting dialogs",
In the context of multimedia meeting recordings and analysis, a new kind of multimedia alignment is presented, which aims at thematically aligning documents with all kind of temporal media. The alignment proposed in this presentation uses the similarities that exist between the documents' content and the speech transcript's content in order to provide temporal indexes to printable documents. Several document content alignment strategies are discussed and evaluated through two particular data sets: press reviews meetings and scientific conference presentations. IM2.DI

Dong Zhang, "Learning influence among human interactions",
Watching a number of players moving along a map, can we tell how many games are there? What are the games? What is the role of each individual player? What are the interactions among players? Starting from these questions, Dong Zhang, a Ph.D. student of IDIAP, describes human interaction modeling using probabilistic graphical models, and a diverse set of applications in the context of meetings, email communication, and multi-player games. IM2.MI

Tobias Kaufmann, "Using Rule-based Knowledge to Improve LVCSR",
We show that an elaborate linguistic model of a natural language can be a valuable knowledge source to improve large vocabulary continuous speech recognition (LVCSR). Our approach is to complement a statistical language model with rule-based linguistic knowledge. A hidden Markov model based recognizer and an N-gram language model are used to compute a word lattice which is subsequently processed by a parser. We observed a statistically significant reduction of the word error rate by favoring word sequences which the parser identified as being grammatically correct. After an overview of the system architecture, the talk will focus on the requirements and the development of the rule-based language model. IM2.SP

Nicolas Moenne-Loccoz, "Interactive Retrieval of Video Sequences from Local Feature Dynamics",
This paper addresses the problem of retrieving video sequences that contain a spatio-temporal pattern queried by a user. To achieve this, the visual content of each video sequence is first decomposed through the analysis of its local feature dynamics. Camera motion of the sequence, background and objects present in the captured scene and events occurring within it are represented respectively by the parameters of the estimated global motion model, the appearance of the extracted local features and their trajectories. At query-time, a probabilistic model of the visual pattern is estimated from the user interaction, captured through a relevance-feedback loop. We show that the method permits to efficiently retrieve video sequences that share, even partially, a spatio-temporal pattern. IM2 .IIR

Mark Barnard, "Event recognition in sports videos using layered HMMs",
The recognition of events in video data is a subject of much current interest. In this presentation, we address several issues related to this topic. The most important of these is over fitting when very large feature spaces are used and relatively small amounts of training data are available. Here we propose a method combining layered HMMs and an unsupervised low level clustering of the features to address this issue of overfitting. Experiments conducted on the recognition task of different events in 7 rugby games demonstrates the potential of our approach with respect to standard HMM techniques coupled with a feature size reduction technique. While the current focus of this work is on events in sports videos, we believe the techniques shown here are general enough to be applied to other sources of data. IM2.MI

Mael Guillemot,"A Meeting Recording Corpus",
In every project related to multimodal interaction, availability of large, common and annotated databases is a critical concern, and an important tool to foster collaboration between the partners. A multi-modal data set consisting 100 hours of meeting recordings will be presented, realized in conjunction with the IM2 and AMI projects. The corpus will eventually be distributed publicly to the scientific community at large. Some of the meetings it contains are naturally occurring, and some are elicited, particularly using a scenario in which the participants play different roles in a design team. The corpus is being recorded using a wide range of devices including close-talking and far-field microphones, individual and room-view video cameras, projection, a whiteboard, and individual pens. It is also being hand-annotated for many different phenomena, including orthographic transcription, discourse properties such as dialogue acts, summaries, emotions, focus of attention and gestures. We will explain explain how the material is being recorded, pre-processed, transcribed, annotated and distributed on the large-capacity (IM)2 media file server. A particular
emphasis of the talk will be given on how can (IM)2 researchers access and benefit from such a corpus. IM2.IP

Siley Ba, "Probabilistic Models for Head Pose Tracking in Meetings" IM2.SA
This talk addresses the problem of head pose estimation in general, with an application to meeting data. More precisely, given a video of people involved in a meeting, the goal is to estimate the pose of people's head with respect to the

camera, which could ultimately be used for the estimation of the people 's focus-of-attention: who is looking at whom or what. To this end, we formulate this task as a joint head tracking and head pose estimation problem in a Bayesian framework.
The solution to this problem involves the definition of the state space, the modelisation of the head appearence likelihood and of the dynamics, and the solving of the recursive filtering equations in a non-linear and non gaussian context. In the talk, we will address these points, by proposing a multimodal head pose modeling (texture+color) with appropriate likelihood normalisation procedure, and by exploring different alternative sampling schemes (importance sampling, Rao-Blackwellization, Markov Chain Monte Carlo) in a particle filter framework to address the filtering issue.
To illustrate the different issues, we conducted different experiments on a publicly available database consisting of people engaged in meeting discussions and for which the groundtruth is available thanks to the use of magnetic field 3D location and orientation tracker (flock-of-birds), as well as on some outdoor sequences to estimate the discrete focus of attention of walking people. IM2.SA

Marita Ailomaa, ”Accessing Multimedia Meeting Data Multimodally”,
This presentation describes experiences gathered during a Wizard of Oz experiment with a multimodal system for multimedia meeting content retrieval and browsing and in particular addresses the problem of eliciting natural language input in such an environment. We analyze which functionalities users associate with which modality and how they express themselves linguistically when using the language modalities. We discuss specific problems with the elicited language data and why it is not representative of the type of language that the speech recognition and language processing capabilities of the system should be built upon. We also discuss what we believe could be done differently in future experiments to overcome these problems. IM2.DM

Wednesday Afternoon

Guest Speaker TBA
or
Pierre W. Ferrez, “Automatic Detection of Interaction Errors from EEG”
Brain-computer interfaces (BCI), as any other interaction modality based on physiological signals and body channels (e.g., muscular activity, speech and gestures), are prone to errors in the recognition of subject's intent. An elegant approach to improve the accuracy of BCIs consists in a verification procedure directly based on the presence of error-related potentials (ErrP) in the EEG recorded right after the occurrence of an error. Most of these studies show the presence of ErrP in typical choice reaction tasks where subjects respond to a stimulus and ErrP arise following errors due to the subject's incorrect motor action. However, in the context of a BCI, the central question is: "Are ErrP also elicited when the error is made by the interface during the recognition of the subject's intent?" In other words, let's imagine that the subject's intent is to make a robot reach a target to the left. What would happen if the interface fails to recognize the intended command and the robot starts turning in the wrong direction? Are ErrP still present even though the subject did not make any error but only perceived that the interface is performing wrongly? We have thus explored whether ErrP also follow a feedback indicating incorrect responses of the interface and no longer errors of the subject himself. Four healthy volunteer subjects participated in a simple human-robot interaction experiment (i.e., bringing the robot to either the left or right side of a room), which seem to reveal a new kind of ErrP. This “Interaction ErrP” exhibits a first sharp negative peak followed by a broader positive peak and a second negative peak (~270, ~400 and ~550 ms after the feedback, respectively). But in order to exploit this Interaction ErrP we need to detect it in each single trial using a short window following the feedback that shows the response of the classifier embedded in the BCI. We have achieved an average recognition rate of correct and erroneous single trials of 86.4% and 77.5%, respectively. These figures have been obtained using a 10-fold cross-validation where testing is always done on a different recording session to those used for training the classifier. We also show that the integration of this Interaction ErrP in a BCI, where the subject’s intent is not executed if an ErrP is detected, yields significant improvements in the bit-rate of the BCI. IM2.MI

Florent Monay/Pedro Quelhas, “Constructing visual models with local descripors and latent aspects”,
In this talk, we present a new approach to build visual models for scenes and objects in image collections, based on the use of local invariant features and probabilistic latent space models. The first step consists in the quantization of local invariant features into a finite set of patterns, to represent images as histograms of such local descriptions. This representation is an important simplification of the original data, as it discards any spatial information between local patterns in an image. We show its discriminative power in object and scene classification tasks.
From these simple histogram representations, we further model the co-occurrence of local patterns in a given image collection using probabilistic latent space analysis (PLSA). This unsupervised learning technique allows to model objects and scenes as mixtures of different visual aspects, and generates a compact and robust representation for scene and object classification. Finally, by exploiting the ability of PLSA to automatically extract visually meaningful aspects, we describe novel algorithms for aspect-based image ranking and context-sensitive image segmentation. IM2 .IIR

Gianluca Monaci, “Analysis of Multimodal Signals Using Redundant Representations”,
In this work we explore the potentialities of a framework for the representation of audio-visual signals using decompositions on overcomplete dictionaries. Redundant decompositions may describe audio-visual sequences in a concise fashion, preserving good representation properties thanks to the use of redundant, well designed, dictionaries. We expect that this will help us overcome two typical problems of multimodal fusion algorithms. On one hand, classical representation techniques, like pixel-based measures (for the video) or Fourier-like transforms (for the audio), take into account only marginally the physics of the problem. On the other hand, the input signals have large dimensionality. The results we obtain by making use of sparse decompositions of audio-visual signals over redundant codebooks are encouraging and show the potentialities of the proposed approach to multimodal signal representation. IM2.SA

Ardhendu Behera, “DocMIR, an Automatic Document-based Indexing System for Meeting Retrieval”,
This presentation describes the DocMIR system that captures, analyzes and then automatically indexes meetings by taking advantage of projected documents. The system requires neither prior preparation nor specific software installed on the speaker's laptop used for the presentation. Indexing is carried out by analyzing the content of the video of the projector screen, in which the system detects the scene changes, extracts the documents and the duration of each projected documents. Further, each of the captured documents is identified from a document repository containing the original electronic documents. The video segments are then enriched with the textual content of the original documents. Finally, meeting retrieval can be done either by querying document images captured from handheld devices or using a set of keywords. The presentation concludes with a full evaluation of the DocMIR system on real data recorded during MLMI 2004. IM2.DI