Poster Session - Abstract list

Wednesday, September 2nd

IM2 Summer Institute 2009
August 31st - September 2nd, 2009
Chavannes-de-Bogis

IM2.AP, Audio Processing			
1	Mutual Information based Channel selection for Speaker Diarization of Meeting Data	*Deepu Vijayasena*	In the meeting case scenario, audio is often recorded using Multiple Distance Microphones (MDM) in a non-intrusive manner. Typically a beamforming is performed in order to obtain a single enhanced signal out of the multiple channels. This paper investigates the use of mutual information for selecting the channel subset that produces the lowest error in a diarization system. Conventional systems perform channel selection on the basis of signal properties such as SNR, cross correlation. In this paper, we propose the use of a mutual information measure that is directly related to the objective function of the diarization system. The proposed algorithms are evaluated on the NIST RT 06 eval dataset. Channel selection improves the speaker error by 1.1% absolute (6.5% relative) w.r.t. the use of all channels.
2	Application of Out-of-Language Detection to Spoken Term Detection	*Petr Motlicek*	This work investigates the detection of spoken terms, performed by a system developed for English, in a conversational multi-language scenario. The speech is processed using a large vocabulary continuous speech recognition system. The recognition output is represented in the form of word recognition lattices which are then used to search required terms. Due to the potential multi-lingual speech segments at the input, the spoken term detection system is combined with a module performing out-of-language detection to adjust its confidence scores. First, experimental results of spoken term detection are provided on the conversational telephone speech database distributed by NIST in 2006. Then, the system is evaluated on a multi-lingual database with and without employment of the out-of-language detection module, where we are only interested in detecting English terms. We evaluate several strategies to combine these two systems in an efficient way.
3	Hierarchical and Parallel Processing of the Modulation Spectrum	*Fabio Valente*	The modulation spectrum is a useful representation of a signal for incorporating dynamic informations. In this work we investigate how to process this information from an ASR perspective. Modulation spectrum frequencies are obtained using a log-scaled filter bank. Parallel and hierarchical approaches are investigated. Parallel processing combines output of independent classifiers trained on different modulation frequencies. Hierarchical processing uses different modulation frequencies at different level of the model. Experiments are run on a meeting transcription LVCSR task and results are reported on the RT05 evaluation data. Hierarchical processing outperforms parallel processing. This model is consistent with several perceptual and physiological studies on auditory processing.
IM2.DMA, Database management and meeting analysis			
4	Multiview Clustering: A Late Fusion Approach Using Latent Models	*Eric Bruno*	Multi-view clustering is an important problem in information retrieval due to the abundance of data offering many perspectives and generating multi-view representations. We investigate in this short note a late fusion approach for multi-view clustering based on the latent modeling of cluster-cluster relationships. We derive a probabilistic multi-view clustering model outperforming an early-fusion approach based on multi-view feature correlation analysis.
5	The ezHub	*Mike Flynn*	The Hub distributes and stores data in real time. The ezHub is a simple, elegant and efficient Java API to produce and consume Hub data, with no need to know about the intricacies of connection, data-formats, error-handling, threading issues, compression, etc.

6	Speaker Change Detection with Privacy-Preserving Audio Cues	*Sree Hari Krishnan Parthasarathi*	In this paper we investigate a set of privacy-sensitive audio features for speaker change detection (SCD) in multiparty conversations. These features are based on three different principles: characterizing the excitation source information using linear prediction residual, characterizing subband spectral information shown to contain speaker information, and characterizing the general shape of the spectrum. Experiments show that the performance of the privacy-sensitive features is comparable or better than that of the state-of-the-art full-band spectral-based features, namely, Mel Frequency Cepstral Coefficients, which suggests that socially acceptable ways of recording conversations in real-life is feasible.
IM2.HMI, Human-machine interaction			
7	In-Meeting Interaction with Printed Paper	*Maurizio Rigamonti*	In this poster we present an on-line meeting assistant providing real-time and natural interaction between the participants and printed papers. The main capabilities of the assistant that we foresee are: 1. Control the slideshows by using printed paper to select the currently projected slide; 2. Add indexes and notes during the meeting; 3. Permit the participant to access additional information related to the content of a slide. The assistant analyzes the video stream captured with a camera over the head of a participant and recognizes her fingers, as well as manipulated documents. Furthermore, the assistant beams augmented information over the paper, e.g. document's structures, interactive graphics, etc., in order to improve user experience. This poster proposes a general overview of the assistant and presents how it detects fingers and documents in real-time video streams.
8	WotanEye: Personal information management through interactive visualizations	*Florian Evéquoz, Denis Lalanne, Rolf Ingold*	Managing efficiently its personal information offers lots of challenges. WotanEye offers a set of interactive visualization techniques allowing to organize and overview one's personal digital memory, and better understand the strategies one can use to manage it. It provides moreover a natural way to explore one's personal information space using facets. Applied to meetings, it is meant to support a meeting participant in gathering heterogeneous pieces of personal information about the meetings he attended or plans to attend. In this poster, we present the challenges of PIM, our approach and an application to IM2 meetings.
9	Computing semantic similarity based on conceptual networks for information retrieval	*Majid Yazdani, Andrei Popescu-Belis*	This work aims at finding new similarity measures between two texts, using human knowledge that is expressed implicitly in semantic networks, in particular networks built from Wikipedia. In the first step, the distance between each two concepts in the network is estimated from a random-walk model over the network. In the second step, a general similarity measure between any two texts is computed using a projection of the documents in the space of the concepts. As concepts are not orthogonal to each other, we generalize the cosine similarity measure for vectors with non orthogonal axes by knowing the distances between axes. The first application of this measure is to a semantic search procedure for just-in-time retrieval over a database of meeting transcripts. The similarity measure defined above is suitable for noisy text retrieval, and allows us to find best timing for sending queries, upon significant semantic changes in the spoken queries.
10	Assessing Multimodal Fusion Engines Performances: Experiments with HephaisTK	*Bruno Dumas, Denis Lalanne, Rolf Ingold*	In interactive multimodal systems, when fusing input data coming from different modalities, the fusion engine and its inner mechanisms play a vital role. Along with ambiguity resolution problems, slight temporal or contextual differences in the way the same information is expressed by the user can completely change the intended meaning. Thus, being able to assess how a given fusion engine architecture behaves in front of problematic use cases shall help the overall user experience. This poster presents our recent experiments on multimodal fusion engines performances assessment with help of the HephaisTK multimodal interfaces creation toolkit.
11	A just-in-time meeting retrieval system using spoken input	*Andrei Popescu-Belis, Alexandre Nanchen, Majid Yazdani, and Mike Flynn*	The Automatic Content Linking Device (ACLD) monitors a conversation using an ASR system, and uses the detected words to retrieve documents that are of potential interest to the participants. The document set includes project related documents such as reports, memos or emails, as well as snippets of past meetings that were transcribed using offline ASR. In addition, results of Web searches are also displayed. The application, which can be demonstrated offline on the AMI/IM2 Meeting Corpus, models the conversational context and refreshes the displayed results accordingly. Feedback from potential users has been positive and has contributed to the improvement of the application.

12	A method for automatic BET question answering in meetings and its evaluation	*Quoc Anh Le and Andrei Popescu-Belis*	Information access in meeting recordings can be assisted by meeting browsers, or can be fully automated following a question-answering (QA) approach. We define an information access task based on the BET, aiming at discriminating true from false parallel statements about facts in meetings. An automatic QA method is then applied to this task, and scores 59% accuracy for passage retrieval (random guess scores below 1%), but only 60% on combined retrieval and question discrimination (baseline is 50% and humans reach 70%–80%). The method greatly outperforms humans for speed, at less than 1 second per question, vs. 1.5–2 minutes. The method is thus a promising enhancement to meeting browsers, as an assistant for relevant passage retrieval.
IM2.MCA, Multimodal context abstraction			
13	Implicit Emotional Tagging of Multimedia Using EEG Signals and Brain Computer Interface	*Ashkan Yazdani*	In multimedia content sharing social networks, tags assigned to content play an important role in search and retrieval. In other words, by annotating multimedia content, users can associate a word or a phrase (tag) with that resource such that it can be searched for efficiently. Implicit tagging refers to assigning tags by observing subjects behavior during consumption of multimedia content. This is an alternative to traditional explicit tagging which requires an explicit action by subjects. In this paper we propose a brain-computer interface (BCI) system based on P300 evoked potential, for implicit emotional tagging of multimedia content. We show that our system can successfully perform implicit emotional tagging and naïve subjects who have not participated in training of the system can also use it efficiently. Moreover, we introduce a subjective metric called “emotional taggability” to analyze the recognition performance of the system, given the degree of ambiguity that exists in terms of emotional values associated with a multimedia content.
14	Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel	*Francesca De Simone*	In this poster we describe a database containing subjective assessment scores relative to 78 video streams encoded with H.264/AVC and corrupted by simulating the transmission over error-prone network. The data has been collected from 40 subjects at the premises of two academic institutions. Our goal is to provide a balanced and comprehensive database to enable reproducible research results in the field of video quality assessment. In order to support research works on Full-Reference, Reduced-Reference and No-Reference video quality assessment algorithms, both the uncompressed files and the H.264/AVC bitstreams of each video sequence have been made publicly available for the research community, together with the subjective results of the performed evaluations.
15	Social Media Tag Propagation	*Ivan Ivanov*	In this poster we present our Social Media Tag Demonstrator (SMTAG) system. It is a web-based tool that allows users to annotate objects in images (put some labels, tags) and then propagate appropriate tags to the same objects found in the database. SMTAG tool is based on object duplicate detection algorithm. It uses database based on Paris landmarks. The Paris dataset consists of images (around 6500) which mainly depict monuments, which are supposed to be unique. The user is free to label as many objects depicted in the image as they choose. We demonstrate current possibilities of SMTAG tool which is still under development.
16	TagCaptcha: Annotating images with CAPTCHAs	*Donn Morrison*	We introduce a method of annotating images for use in a retrieval setting by exploiting the need for CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) online. Our system, called TagCaptcha, presents the user with a number of images that must be correctly labelled in order to pass the test. The images are divided into two subsets: a control or verification set for which annotations are known, and an unknown set for which no verified annotations exist. The verification set is used to control against the tags provided for the unknown set. If the user provides correct verification tags, the tags for the unknown set are promoted. An image with a promoted tag must be validated by other users before it can be classed as annotated and added to the verification set. Given a partially annotated database, the images can be incrementally annotated over time. We report usability results from a small user study as well as sample user tags from the online demonstration system.

IM2.MPR, Multimodal Processing and recognition			
17	Audio-visual speaker diarization and localization	*Hayley Hung*	We present speaker diarization and localization work which has been carried out between Idiap and ICSI for the last year. This work has approached the speaker diarization and localization problem through many different practical set-ups as well as in off-line and on-line scenarios. We present the evolution of our research over this period culminating finally in a single camera single microphone set-up. Evaluations have been made both in terms of the speaker diarization error but also localization error, showing both the benefits of fusing audio and video features for improved performance as well as exploiting general body motion as a cue for speaking.
18	Verification of aging face using stacking method	*Weifeng Li, Andrzej Drygajlo*	Permanence of biometric features for face verification remains a largely open research problem. Actual and up-to-date at the time of their creation, extracted features and models relevant to a person's face may eventually become outdated, leading to a failure in the face verification task. If physical characteristics of the individual change over time, their classification model has to be updated. In this paper, we develop a Q-stack classifier that performs face verification across age progression. Originally, Q-stack classifier has been proposed to use class-independent signal quality measures and baseline classifier scores in order to improve classification. In this paper we demonstrate the application of Q-stack classifier on the task of biometric identity verification using face images and associated metadata quality measure - age. We show that the use of the proposed technique allows for reducing the error rates below those of baseline classifier created at the time of enrolment.
19	Haar Local Binary Pattern Feature for Fast Illumination Invariant \\Face Detection	*Anindya Roy*	Face detection is the first step in many visual processing systems like face recognition, emotion recognition and lip reading. In this paper, we propose a novel feature called Haar Local Binary Pattern (HLBP) feature for fast and reliable face detection, particularly in adverse imaging conditions. This binary feature compares bin values of Local Binary Pattern histograms calculated over two adjacent image subregions. These subregions are similar to those in the Haar masks, hence the name of the feature. They capture the region-specific variations of local texture patterns and are boosted using AdaBoost in a framework similar to that proposed by Viola and Jones. Preliminary results obtained on several standard databases show that it competes well with other face detection systems, especially in adverse illumination conditions.
20	Asynchronous models for audio-visual speech recognition	*Virginia Estellers, Mihai Gurban, Jean-Philippe Thiran*	We investigate the use of asynchronous statistical models for audio-visual speech recognition. The models integrating the audio and visual data are based on dynamic bayesian networks and multistream HMMs, with different levels of asynchrony and stream independence possible. We propose a new model correctly exploiting the asynchrony and compare its performance to traditional ones with different synchrony and dependency constraints. A first set of experiments suggests the necessity of processing the data streams in order to reduce their asynchrony and allow simpler statistical models. We develop such a processing method adapting the video to the time evolution of the audio stream and test its effects on a typical audio-visual speech recognition system.
21	Audio-Visual Processing of Video Sequences for Visual Attention Analysis.	*Basilio Noris, Johnny Mariéthoz, François Fleuret, Aude Billard*	We present in this work the multi-modal processing of a video stream acquired with a wearable camera. The main objective of this processing is to provide automatic measurements related to visual attention in response to directed speech. The system was evaluated in a settings with two external speakers and a quiet observer wearing the camera. Visual processing consists of object detection and localization in the scene, coupled with a gaze tracker. It allows to determine the attentional focus of the individual wearing the camera. Audio processing detects automatically a series of keywords in free speech uttered by one of the two external speakers. The combination of these two sources of information allows to correlate the utterance of certain words ("ball", "pen", etc.) with visual attention shift towards the corresponding objects in the observer. Quantitative validation of the performance of the system for visual and auditory processing are presented separately.

IM2.VP, Visual/Video Processing			
22	Volterra Series for Analyzing MLP based Phoneme Posterior Estimator	*Joel Pinto, G. Sivaram, Hynek Hermanksy and Mathew Magimai.-Doss*	We present a framework to apply Volterra series to analyze multilayered perceptrons trained to estimate the posterior probabilities of phonemes in automatic speech recognition. The identified Volterra kernels reveal the spectro-temporal patterns that are learned by the trained system for each phoneme. To demonstrate the applicability of Volterra series, we analyze a multilayered perceptron trained using Mel filter bank energy features and analyze its first order Volterra kernels.
23	Hough transform-based mouth localization for audio-visual speech recognition	*Gabriele Fanelli, Juergen Gall*	We present a novel method for mouth localization in the context of multimodal speech recognition where audio and visual cues are fused to improve the speech recognition accuracy. While facial feature points like mouth corners or lip contours are commonly used to estimate at least scale, position, and orientation of the mouth, we propose a Hough transform-based method. Instead of relying on a predefined sparse subset of mouth features, it casts probabilistic votes for the mouth center from several patches in the neighborhood and accumulates the votes in a Hough image. This makes the localization more robust as it does not rely on the detection of a single feature. In addition, we exploit the different shape properties of eyes and mouth in order to localize the mouth more efficiently. Using the rotation invariant representation of the iris, scale and orientation can be efficiently inferred from the localized eye positions. The superior accuracy of our method and quantitative improvements for audio-visual speech recognition over monomodal approaches are demonstrated on two datasets.
24	Dynamic Partitioned Sampling For Tracking With Discriminative Features	*Stefan Duffner, Jean-Marc Odobez, Elisa Ricci*	We present a multi-cue fusion method for tracking with particle filters which relies on a novel hierarchical sampling strategy. Similarly to previous works, it tackles the problem of tracking in a relatively high-dimensional state space by dividing such a space into partitions, each one corresponding to a single cue, and sampling from them in a hierarchical manner. However, unlike other approaches, the order of partitions is not fixed a priori but changes dynamically depending on the reliability of each cue, i.e. more reliable cues are sampled first. We call this approach Dynamic Partitioned Sampling (DPS). The reliability of each cue is measured in terms of its ability to discriminate the object with respect to the background, where the background is not described by a fixed model or by random patches but is represented by a set of informative "background particles" which are tracked in order to be as similar as possible to the object. The effectiveness of this general framework is demonstrated on the specific problem of head tracking with three different cues: colour, edge and contours. Experimental results prove the robustness of our algorithm in several challenging video sequences.
25	Text non-Text Classification of Handwritten Document Zones	*Emanuel Indermühle, Horst Bunke*	Two systems distinguishing text and non-text content in handwritten documents are presented in this poster. In the first approach, documents are segmented into zones using layout analysis. A state-of-the art classifier distinguishes between text and non-text segments. Under this classifier, four different segmentation methods (X-Y cut, morphological closing, Voronoi segmentation, and whitespace analysis) are compared. The second system, which follows a bottom-up approach, classifies connected components. With the best method up to 94.3% of the pixels can be correctly assigned. The experiments are performed on a new large dataset of online handwritten documents containing different content types in arbitrary arrangement.
IM2 OTHERS			
26	OCD & Dolores: A Complete and Dynamic Document Physical and Logical Restructuring Workflow	*Jean-Luc Bloechle*	Dolores is an innovative system for recovering the logical structure of electronic documents. Dolores uses XED's output as input, i.e., OCD files containing physical structures extracted from PDF documents. The originality of our work consists firstly in its pivot format, i.e., OCD, and secondly in its flexible and efficient approach: document structures are interactively learned, the inferred knowledge is instantly injected in the system and reflected to the user through a dynamic graphical interface.

27	Multimodal relevance feedback for image retrieval	*Nicolae Suditu*	For collections where each image is accompanied by a text (eg. COREL database of annotated photos), we propose a multimodal relevance feedback approach able to take advantage of both visual and textual information. The users are guided towards the desired image or image category through several feedback iterations in which they are asked to show, among a small set of images, the image that best matches what they search for. At each iteration, the retrieval algorithm reestimates the probability of relevance for all images in the collection and selects consequently the next set of images seeking to maximize the feedback information gain. In this process, visual and textual features are dynamically weighted based on the users feedback. Starting from a random set of images, no query is required and this is a major advantage as the visual content of images is often difficult to describe in terms of keywords.
28	Flickr Hypergroups	*Radu-Andrei Negoescu, Brett Adams, Dinh Phung, Sveta Venaktesh, Daniel Gatica-Perez*	The amount of multimedia content available online constantly increases, and this leads to problems for users who search for content or similar communities. Users in Flickr often self-organize in user communities through Flickr Groups. These groups are particularly interesting as they are a natural instantiation of the content~+~relations social media paradigm. We propose a novel approach to group searching through hypergroup discovery. Starting from roughly 11,000 Flickr groups' content and membership information, we create three different bag-of-word representations for groups, on which we learn probabilistic topic models. Finally, we cast the hypergroup discovery as a clustering problem that is solved via probabilistic affinity propagation. We show that hypergroups so found are generally consistent and can be described through topic-based and similarity-based measures. Our proposed solution could be relatively easy implemented as an application to enrich Flickr's traditional group search.
29	Topic Models for Scene Analysis and Abnormality Detection	*Jagannadan Varadarajan, Jean-Marc Odobez*	Automatic analysis and understanding of common activities and the detection of deviant behaviors is a challenging task in computer vision. This is particularly true in surveillance data, where busy traffic scenes are rich with multifarious activities many of them occurring simultaneously. In this paper, we address these issues and show how an unsupervised learning approach relying on probabilistic Latent Semantic Analysis (pLSA) applied to a rich set of visual features including motion and size activities allows to discover relevant activity patterns occurring in such scenes. We show how the discovered patterns can directly be used to segment the scene into regions with clear semantic interpretation. Furthermore, we introduce novel abnormality detection measures within the scope of the adopted modeling approach, and investigate in detail their performance with respect to various issues. Experiments on 45 minutes of video captured from a busy traffic scene and involving abnormal events are conducted.