Poster Session - Abstract list

Tuesday, September 14th

IP1: Integrated Multimodal Processing

1	Kernel Weighing using the Margin	*Basilio Noris*	We study the problem of kernel weighing in a SVM one-vs-many classification task by exploring the boundary between classes. A standard SVM is used to identify training samples inside the classification margin, and kernel weighing techniques are applied to extract relevant features for an object detection/classification task. We then give an incremental scheme for training a classifier on a clean dataset and adapting it to a new noisier dataset with fewer samples.
2	Object-based Tag Propagation for Semi-Automatic Annotation of Images	*Ivan Ivanov*	Over the last few years, social network systems have greatly increased users' involvement in online content creation and annotation. Since such systems usually need to deal with a large amount of multimedia data, it becomes desirable to realize an interactive service that minimizes tedious and time-consuming manual annotation. We propose an interactive online platform that is capable of performing semi-automatic image annotation and tag recommendation for an extensive online database. First, when the user marks a specific object in an image, the system performs an object duplicate detection and returns the search results with images containing similar objects. Then, the annotation of the object can be performed in two ways: (1) In the tag recommendation process, the system recommends tags associated with the object in images of the search results, among which, the user can accept some tags for the object in the given image. (2) In the tag propagation process, when the user enters his/her tag for the object, it is propagated to images in the search results. Different techniques to speed-up the process of indexing and retrieval are presented and their effectiveness demonstrated through a set of experiments considering various classes of objects.
3	Subjective Evaluation of Next-Generation Video Compression Algorithms: a Case Study	*Francesca De Simone*	In this poster, the details and the results of the subjective quality evaluation performed at EPFL, as a contribution to the Joint Collaborative Team on Video Coding (JCT-VC) High Efficiency Video Coding (HEVC) effort for the definition of the next generation video coding standard, are described. The performance of twenty-seven coding technologies have been evaluated, with respect to two H.264/MPEG AVC anchors, on High Denition (HD) test material, i.e. HD1080p up to 60 fps and HD720p at 60 fps, as defined in the Call for Proposal issued by JCT-VC in January 2010. The test campaign involved a total of 494 naive observers, over a period of four weeks. Similar efforts have been carried out as part of the standardization process of previous video coding technologies, but the test campaign described in this paper was by far the most extensive one in the history of video coding standardization. The obtained results show high consistency and allow an accurate comparison of codec performance.
4	Gesture and touch controlled video player interface for mobile devices	*Francesca De Simone*	Today mobile communication devices allow users to access a wide variety of multimedia contents and services. In order to improve user experience and device usability, the design of interfaces and interaction techniques for mobile devices have focused on new modalities, other than those used for desktop computers. In this paper, we describe a novel gesture controlled video player interface for mobile devices. The results of a usability study confirm that users would definitely like to adopt the major part of the proposed features. Furthermore, the responsiveness and reliability of the interface has been studied. Measured response times have been found to be within acceptable boundaries and the number of unrecognized haptic controls is limited.

5	Discriminating group interaction types using nonverbal conversational behavior	*Dinesh Babu Jayagopi*	In this poster, we describe our approach to discriminating group interaction types using nonverbal conversational behavior. We motivate the problem, our approach and finally show the results on two problems - our recent work on discriminating brainstorming and decision-making meetings and our previous work on discriminating cooperative and competitive meetings.
6	Epitome - A Social Game for Photo Album Summarization	*Ivan Ivanov, Peter Vajda, Jong-Seok Lee, Touradj Ebrahimi*	With the rapid growth of digital photography, sharing of photos with friends and family has become very popular. When people share their photos, they usually organize them in albums according to events or places. To tell the story of some important events in one's life, it is desirable to have an efficient summarization tool which can help people to get a quick overview of an album containing huge number of photos. We propose an approach for photo album summarization through a novel social game "Epitome" for mobile phones. Our approach to album summarization consists of two games: "Select the Best!" and "Split it!". The goal of the first game is to allow a user to select the most representative photo of a reduced set of images, while in the second game, the user has to split the reduced set into two distinct parts. As it could be time-consuming to look at a huge collection of photos on a mobile phone, it is more enjoyable and pleasant to show only a limited number of images which can be fit into one mobile screen. The results obtained in these games are combined to produce a summarization and are then compared with the results of other users. As a final result, a unique summarization sequence of photos is determined. The determined sequence of photos can be used to create a collage of one album or a cover for an album. The proof of concept of the proposed method is demonstrated through a set of experiments on several photo collections.
7	Mobile Visual Search	*Peter Vajda, Ivan Ivanov, Touradj Ebrahimi*	Over the last few years, the mobile market is growing rapidly and several mobile application using content retrieval algorithms were born. In this poster, we propose an interactive mobile application for object based retrieval. First, users take photos from different objects and describe them with text or web page link. The information is sent to a server, where the photos are processed. Later, when a user takes a picture from a learnt object, he will receive a list of information from the database that contains contributions of several users. Different techniques to speed up the process of indexing and retrieval are presented in this poster and their effectiveness is demonstrated through a set of experiments considering various classes of objects.
8	Parallel Cross Modal Search Engine	*Rihui Chen and Marc von Wyl*	When there is a huge amount of records in a single database, even with fast algorithms a search process can become unacceptably slow. We propose a parallel version of a multimodal search engine where the search process is parallelized by using a distributed database system and message passing interface for communication. Multiple machines in a cluster are used to do the data retrieving and computation simultaneously so that the time needed for a search query is reduced significantly and the search process can be achieved in an acceptable time.
9	Voices of Vlogging	*Joan-Isaac Biel & Daniel Gatica-Perez*	Vlogs have rapidly evolved from the 'chat from your bedroom' format to a highly creative form of expression and communication. However, despite the high popularity of vlogging, automatic analysis of conversational vlogs have not been attempted in the literature. In this paper, we present a novel analysis of conversational vlogs based on the characterization of vloggers' nonverbal behavior. We investigate the use of four nonverbal cues extracted automatically from the audio channel to measure the behavior of vloggers and explore the relation to their degree of popularity and that of their videos. Our study is validated on over 2200 videos and 150 hours of data, and shows that one nonverbal cue (speaking time) is correlated with levels of popularity with a medium size effect.

10	***Demonstration*:** Demonstration of a Speech-based Just-in-Time Retrieval System for Multimedia Documents	*Andrei Popescu-Belis and Alexandre Nanchen*	The Automatic Content Linking Device (ACLD) is a just-in-time retrieval system that monitors an ongoing conversation or a monologue and enriches it with potentially related documents, including transcripts of past meetings, from local repositories or from the Internet. The linked content is displayed in real-time to the speakers. The system will be demonstrated in the single-speaker real-time setting, using real-time automatic speech recognition, with a flexible user interface that displays labels of results and provides access to their content.
IP2: Human Centered Design & Evaluation			
11	Computing semantic distance using contents and hyperlinks of Wikipedia articles	*Majid Yazdani , Andrei Popescu-Belis*	We propose a method for computing textual semantic similarity by using knowledge from Wikipedia, namely the contents of articles and the hyperlinks between them. A network of concepts is built from corresponding Wikipedia documents, with two types of weighted links, based respectively on actual hyperlinks and on the lexical similarity between articles. We propose and implement an efficient random walk algorithm that computes the visiting probability from one set of nodes to another, as distance between sets of nodes. To evaluate the proposed distance, we first show that our system has concept clustering abilities that are comparable to human ones, over ten selected subsets of documents from the English Wikipedia. Second, our system was applied to two benchmark semantic tasks: word similarity and document similarity. The results show that our system reaches state-of-the-art performance on each task, and that using both hyperlinks and lexical similarity links improves the scores over using only one of them.
12	Augmented Teams Specification: Applying Content Linking in an Educational Setting	*Nan Li, Frédéric Kaplan, Pierre Dillenbourg, Majid Yazdani, Alexandre Nanchen, Andrei Popescu-Belis*	This poster introduces the specification of the Augmented Teams application and discusses the challenges related to porting the Automatic Content Linking Device to the EPFL Rolex Learning Center, and proposes solutions for a tabletop interface. The system is intended to monitor discussions in the collaborative context of a student working room. The system will (1) display a representation of the important words that are automatically recognized, (2) allow participants to act on these words and use them to query various repositories, and (3) display results and help participants to use them.
13	Emotion recognition using Brain Computer Interface	*Ashkan Yazdani, Jong-Seok Lee, Touradj Ebrahimi*	Recently, the field of automatic recognition of users' affective states has gained a great deal of traction. Automatic, implicit recognition of affective states has many applications, ranging from personalized content recommendation to automatic tutoring systems. In this work, we present an early, promising approach to classification of emotions induced by watching music videos. We show robust correlations between users' self-assessments of arousal and valence and the frequency powers of their EEG activity. We present methods for single trial classification using both EEG and peripheral physiological signals.
14	Recognition of hand gestures for a novel economic HCI	*Matthias Schwaller, Denis Lalanne*	This poster presents a PhD work on hand gesture recognition to develop a novel economic HCI, avoiding non ergonomic and tiring movements. The scenario in which this novel way to interact will be integrated is the Communication Board (CBoard), in which several users can collaborate remotely like if they were at the same place. This PhD thesis work focuses on pointing and selecting, with visual feedbacks to augment usability and precision. The poster presents the current work achieved in this direction and in particular the Portable Gestural Interface PyGmI, which we implemented to visualize and navigate into presentation files, thanks to a tiny projector fixed on the user's belt. This poster presents PyGmI, its setup, the designed gestures, the recognition modules, an application using it and finally an evaluation.

IP3: Social Signal Processing			
15	The voice of personality: Mapping non-verbal vocal behavior into trait attributions	*Vinciarelli Alessandro and Gelareh Mohammadi*	This paper reports preliminary experiments on automatic personality traits attribution based on non-verbal vocal behavioral cues.
16	Automatic Speaker Segmentation of Political Debates using Role based Turn-Taking Patterns	*Fabio Valente and Alessandro Vinciarelli*	Several recent works on social signals have addressed the problem of statistical modeling of social interaction in multi-party discussions showing that characteristics like turn-taking patterns can be modeled and predicted according to the role that each participant has in the discussion. Reversely this work investigates the use of social signals to improve conventional speech processing methods. In details we propose the use of turn-taking patterns induced by roles for improving speaker diarization, the task of determining 'Who spoke when' in an audio file.
17	Multistream Speaker Diarization beyond Two Acoustic Feature Streams	*Deepu VIJAYASENAN*	Speaker diarization for meetings data are recently converging towards multistream systems. The most common complementary features used in combination with MFCC are Time Delay of Arrival (TDOA). Also other features have been proposed although, there are no reported improvements on top of MFCC+TDOA systems. In this work we investigate the combination of other feature sets along with MFCC+TDOA. We discuss issues and problems related to the weighting of four different streams proposing a solution based on a smoothed version of the speaker error. Experiments are presented on NIST RT06 meeting diarization evaluation. Results reveal that the combination of four acoustic feature streams results in a 30% relative improvement with respect to the MFCC+TDOA feature combination. To the authors' best knowledge, this is the first successful attempt to improve the MFCC+TDOA baseline including other feature streams.
IM2 OTHERS			
18	Hierarchical Multilayer Perceptron based Language Identification	*David Imseng*	Automatic language identification (LID) systems generally exploit acoustic knowledge, possibly enriched by explicit language specific phonotactic or lexical constraints. This paper investigates a new LID approach based on hierarchical multilayer perceptron (MLP) classifiers, where the first layer is a "universal phoneme set MLP classifier". The resulting (multilingual) phoneme posterior sequence is fed into a second MLP taking a larger temporal context into account. The second MLP can learn/exploit implicitly different types of patterns/information such as confusion between phonemes and/or phonotactics for LID. We investigate the viability of the proposed approach by comparing it against two standard approaches which use phonotactic and lexical constraints with the universal phoneme set MLP classifier as emission probability estimator. On SpeechDat(II) datasets of five European languages, the proposed approach yields significantly better performance compared to the two standard approaches.
19	Implementation of VTLN for Statistical Speech Synthesis	*Lakshmi Saheer*	Vocal tract length normalization (VTLN) is an important feature normalization technique that can be used to perform speaker adaptation when very little adaptation data is available. It was shown earlier that VTLN can be applied to statistical speech synthesis and was shown to give additive improvements to model based adaptations. This paper presents an expectation maximization (EM) optimization for estimating more accurate warping factors. The EM formulation helps to embed the feature normalization in the hidden Markov model training. This helps in estimating the warping factors more efficiently and enables the use of multiple (appropriate) warping factors for different state clusters of the same speaker.

20	A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-Based Speech Synthesis	*Hui Liang*	The EMIME project aims to build a personalized speech-to-speech translator, such that spoken input of a user in one language is used to produce spoken output that still sounds like the user's voice however in another language. This distinctiveness makes unsupervised cross-lingual speaker adaptation one key to the project's success. So far, research has been conducted into unsupervised and cross-lingual cases separately by means of decision tree marginalization and HMM state mapping respectively. In this paper we combine the two techniques to perform unsupervised cross-lingual speaker adaptation. The performance of eight speaker adaptation systems (supervised vs. unsupervised, intra-lingual vs. cross-lingual) is compared using objective and subjective evaluations. Experimental results show the performance of unsupervised cross-lingual speaker adaptation is comparable to that of the supervised case in terms of spectrum adaptation in the EMIME scenario, even though automatically obtained transcriptions have a very high phoneme error rate.
21	On joint Modelling of Grapheme and Phoneme Information using KL-HMM for ASR	*Ramya Rasipuram*	State-of-the-art automatic speech recognition (ASR) system typically use phoneme as sub-word unit. In this poster, we present a novel approach to jointly model phoneme and grapheme information using Kullback- Deibler divergence-based hidden Markov model (KL-HMM) system. More specifically, the underlying sub-word unit models are represented by graphemes (thus making the dictionary compilation easy), and the phonetic information is modeled/captured through phoneme posterior features estimated using a multilayer perception. We investigated this novel approach by conducting ASR studies on DARPA Resource Management corpus. In particular, we explored the effect of grapheme sub-word context modeling on the performance of the system. Our studies show that through modeling of context information using early tagged sub-word units, grapheme-based ASR system (4.7% WER) could achieve similar performance as state-of-the-art phoneme-based ASR system (4.6% WER).
22	MLP-based Posterior Features for Principled Template-based Speech Recognition	*Serena Soldo*	This paper further investigates the use and robustness of MLP-based posterior features in the context of isolated word recognition, using new type of template based approach. An advantage of the proposed template based approach is that it could generalize to subword-based speech recognizer. One particularly interesting instance of our method is when the test templates/sequences are phone posterior sequences while the reference words are also templates/sequences of multinomial distributions. The same framework can then be used in the context of different local matching “scores” between posterior distributions, including weighted symmetric KL-divergence, Bhattacharya distance, cosine angle, dot product, and L1-norm. Results on task and speaker independent Phonebook data (for both 75 and 600 words lexicon) show that using an MLP trained on an independent “auxiliary” data set can yield results comparable to the use of an MLP trained on matched condition and to state-of-the-art HMMs.
23	Sparse Component Analysis for Robust Speech Recognition	*Afsaneh Asaei, Herve Bourlard, Philip N. Garner*	Sparse Component Analysis is a relatively young technique that relies upon a representation of signal occupying only a small part of a larger space. Mixtures of sparse components are disjoint in that space. As a particular application of sparsity of speech signals, we investigate the DUET blind source separation algorithm in the context of speech recognition for multi-party recordings. We show how DUET can be tuned to the particular case of speech recognition with interfering sources, and evaluate the limits of performance as the number of sources increases. We show that the separated speech fits a common metric for sparsity, and conclude that sparsity assumptions lead to good performance in speech separation and hence ought to benefit other aspects of the speech recognition chain.