Poster Session - Abstract list

Friday, September 2nd

Idiap 20th anniversary
IM2 Summer Institute
September 1st - 2nd, 2011
Idiap Research Institute, Martigny

IP1: Integrated Multimodal Processing

1	Head Pose Detection Using Fast Robust PCA for Side Active Appearance Models Under Occlusion	Anil Yuce, Matteo Sorci, Jean-Philippe Thiran	Face detection and the numerous applications it leads to are now a part of our everyday lives and can be found in any electronic device. Detecting a face and its facial features in uncontrolled environments, however, come along with two main problems that we address in this paper: handling large variations in pose and facial occlusions. The Active appearance model (AAM) is a very efficient method to model and describe the face, but it is not robust against neither of these problems. In this paper we propose to use different AAMs for different pose configurations and introduce a switching method that makes use of the Fast Robust PCA reconstruction technique to decide on the final model to use. The presented method is a very simple and efficient one that is also robust against occlusions. We give results of experiments performed on a database of artificially occluded images to prove that the method is highly effective in detecting the pose of faces even in presence of occlusions.
2	Swiss Cheese for Visual Search	Ivan Ivanov, Peter Vajda, Touradj Ebrahimi	We present Cheese (http://cheese.epfl.ch/), an advanced image management platform for online use and mobile devices. Beside standard features such as image upload, tagging and keyword based search, it offers the user visual similarity based search, object based tagging and semi-automatic tag propagation. For improved interoperability between different image repositories and applications, the platform supports the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format.
3	Recognition of Hand Gestures for a Novel Economic HCI	Matthias Schwaller	This PhD work aims at developing a novel economic interaction paradigm based on hand gestures, 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. The PhD thesis currently focuses on deictic gestures (pointing and selection), with visual feedbacks to augment usability and precision. The poster presents the works achieved in this direction and the results achieved with a time of flight camera.
4	Stochastic Unfolding	Ke Sun	High dimensional embedding techniques have been a thriving research topic in the recent decade and have great potential in multimedia processing. This work proposes a new embedding algorithm called Stochastic Unfolding. It glues nearby points together and flatten the manifold with a stochastic objective function. It is closely related to three families of manifold learning techniques including Stochastic Neighbor Embedding, Laplacian Eigenmaps and Maximum Variance Unfolding. It provides new insight in the high dimensional embedding techniques.

5	Large-Scale Multimedia Retrieval: Distributing Multimodal Interactive Learning	Hisham Mohamed and Stephane Marchand-Maillet	Multimedia retrieval may be achieved based on user feedback interpretation. Machine learning strategies can be designed to help in performing information fusion to gather and exploit every piece of knowledge the user is providing to the system. The scalability of multimedia retrieval becomes a very critical problem, due to the vast increase in the volume of multimedia data. Parallel computing represented in its distributed and shared memory architectures is a solution to design and implement volume-scalable solutions for interactive retrieval within large collections of items bearing multimodal information. In this poster, we emphasise and review both aspects of retrieval performance and robustness against the increase in the scale of the database and in the complexity of the data, based on parallel and distributed algorithms. We summarise our achievements that have already results into concrete developments.
6	Aggregation of Asynchronous Eye-Tracking Streams from Sets of Uncalibrated Panoramic Images	Basilio Noris	Obtaining statistics on wearable eye-tracking data from multiple subjects is very challenging. Indeed each recording provides eye-tracking information with respect to the position and direction of the person, which is not constant across subjects. We provide a means for aggregating these asynchronous data by providing local tracking of the person's field of view (FoV) on panoramic images of the environment. We use scale invariant descriptors to find matches between the reference panoramic image and the eye-tracked FoV image. As the panoramic image is heavily distorted, standard methods for homography fail to compute extrinsic camera parameters for the FoV image. We propose to compute these parameters through the globally-convergent method-of-moving-asymptotes. Our methods succeeds in correctly aligning the FoV-image streams from 40 participants in a real-world setting.
7	Computing Text Semantic Relatedness using the Contents and Links of a Hypertext Encyclopedia	Majid Yazdani, Andrei Popescu-Belis	We propose a method for computing semantic relatedness between words or texts by using knowledge from hypertext encyclopedias such as Wikipedia. A network of concepts is built by filtering the encyclopedia's articles, each concept corresponding to an article. Two types of weighted links between concepts are considered: one based on hyperlinks between the texts of the articles, and another one based on the lexical similarity between them. We propose and implement an efficient random walk algorithm that computes the distance between nodes, and then between sets of nodes, using the visiting probability from one (set of) node(s) to another. Moreover, to make the algorithm tractable, we propose and validate empirically two truncation methods. To evaluate the proposed distance, we apply our method to four important tasks in natural language processing: word similarity, document similarity, document clustering, and information retrieval. The performance of the method is state-of-the-art or close to it for each task, thus demonstrating the generality of the knowledge resource. Moreover, using both hyperlinks and lexical similarity links improves the scores over using only one of them, because hyperlinks bring additional real-world knowledge not captured by lexical similarity.
8	Let Epitome summarize your photo collection!	Ivan Ivanov, Peter Vajda, Touradj Ebrahimi	We present a novel social game "Epitome" for photo album summarization as an Android and Facebook application (http://apps.facebook.com/epitome/). "Epitome" is a social application, which provides many pleasant hours while playing it and enjoying photos. At the same time, it summarizes photo albums and provides useful research data. Users play with photos of their Facebook friends through two games. In these games, the user has to select either better of two photos or pair of photos that is more different. Results of these two games are integrated to produce a summarization for a Facebook photo album. These photos can be used to create a collage of an album, a cover for an album, or to be included in a photo book.

9	Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation	Hui Liang, John Dines	Within the HMM state mapping-based cross-lingual speaker adaptation framework, the minimum Kullback-Leibler divergence criterion has been typically employed to measure the similarity of two average voice state distributions from two respective languages for state mapping construction. Considering that this simple criterion doesn't take any language-specific information into account, we propose a data-driven, phonological knowledge guided approach to strengthen the mapping construction - state distributions from the two languages are clustered according to broad phonetic categories using decision trees and mapping rules are constructed only within each of the clusters. Objective evaluation of our proposed approach demonstrates reduction of mel-cepstral distortion and that mapping rules derived from a single training speaker generalize to other speakers, with subtle improvement being detected during subjective listening tests.
10	On Combining Acoustic and LVCSR based Keyword Spotting Systems	Petr Motlicek	This poster investigates detection of English keywords in a conversational scenario using a combination of acoustic and LVCSR based keyword spotting systems. Acoustic KWS systems search predefined words in parameterized spoken data. Corresponding confidences are represented by likelihood ratios given the keyword models and a background model. First, due to the especially high number of false-alarms, the acoustic KWS system is augmented with confidence measures estimated from corresponding LVCSR lattices. Then, various strategies to combine scores estimated by acoustic and several LVCSR based KWS systems are explored. We show that a linear regression based combination significantly outperforms other (non-linear) techniques. Due to that, the relative number of false-alarms of the combined KWS system decreased by more than 50% compared to the acoustic KWS system. Finally, attention is also paid to the complexities of the KWS systems enabling them to potentially be exploited in real detection tasks.
IP2: Human Centered Design & Evaluation			
11	The Communication Board	Denis Lalanne, Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, Dinesh Babu Jayagopi, Daniel Gatica-Perez	The Communication Board (CBoard) is a vertical surface on which multiple users can interact (using 3D pointing systems) at the same time in co-presence or remotely. The CBoard application serves several goals; it is an applicative framework that uses IM2 technologies to analyze users interactions (speech features currently) and elicit factors influencing collaboration, both to facilitate user evaluations (with automatic low level analysis) and to set up real time technologies (to be integrated in the CBoard). The poster presents the work achieved on the CBoard, its set up, and the result of an evaluation involving 40 teams of 3 users.
12	AugmentedTeams: A Tabletop Environment for Augmenting Meetings with Background Search	Nan Li, Omar Mubin, Frédéric Kaplan, Pierre Dillenbourg , Andrei Popescu-Belis	Web searches are often needed in meetings. Many research projects have been conducted for supporting collaborative search in information-seeking tasks, where the searches are executed both intentionally and intensively. However, for most realistic meetings, Web searches happen randomly with low-intensity. They neither serve as main tasks nor major activities in a meeting. This kind of search is called background search. We propose AugmentedTeams, an augmented tabletop environment with a semi-ambient conversation-context-aware surface as well as foldable paper browsers for facilitating meetings with background search.

IP3: Social Signal Processing			
13	Multistream Speaker Diarization Beyond Two Acoustic Feature Streams	Sree Harsha Yella, Ashtosh Sapru, Fabio Valente and 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.
14	Does Human Action Recognition Benefit from Pose Estimation?	Angela Yao, Juergen Gall, Gabriele Fanelli, Luc Van Gool	The earliest works on human action recognition were focused on tracking and classifying articulated motions of the body. However, they required accurate tracking of body parts, which is a difficult task, particularly under realistic imaging conditions. As such, recent trends have shifted towards the use of more abstract and low-level appearance features such as spatio-temporal interest points. Given the recent progress in pose estimation, we feel that pose-based action recognition systems are now feasible and warrant a second look. In this paper, we address the question of whether pose estimation is useful for action recognition or if it is better to train a classifier only on low-level appearance features drawn from video data. We compare pose-based, appearance-based and combined pose and appearance features for action recognition in a home-monitoring scenario. Our experiments show that pose-based features outperform low-level appearance features, even when heavily corrupted by noise, leading us to the conclusion that pose estimation may be beneficial for the action recognition task.
15	Joint Data Association and Grouping	Stefano Pellegrini, Andreas Ess, Luc Van Gool	We consider the problem of data association in a multi-person tracking context. In semi-crowded environments, people are still discernible as individually moving entities, that undergo many interactions with other people in their direct surrounding. Finding the correct association is therefore difficult, but higher-order social factors, such as group membership, are expected to ease the problem. However, estimating group membership is a chicken-and-egg problem: knowing pedestrian trajectories, it is rather easy to find out possible groupings in the data, but in crowded scenes, it is often difficult to estimate closely interacting trajectories without further knowledge about groups. To this end, we propose a third-order graphical model that is able to jointly estimate correct trajectories and group memberships over a short time window. A set of experiments on challenging data underline the importance of joint reasoning for data association in crowded scenarios.
16	An Integrated Framework for Multi-Channel Multi-Source Localization and Voice Activity Detection.	Mohammad J. Taghizadeh, Philip N. Garner and Herve Bourlard	Two of the major challenges in microphone array based adaptive beamforming, speech enhancement and distant speech recognition, are robust and accurate source localization and voice activity detection. This paper introduces a spatial gradient steered response power using the phase transform (SRP-PHAT) method which is capable of localization of competing speakers in overlapping conditions. We further investigate the behavior of the SRP function and characterize theoretically a fixed point in its search space for the diffuse noise field. We call this fixed point the \textit{null} position in the SRP search space. Building on this evidence, we propose a technique for multi-channel voice activity detection (MVAD) based on detection of a maximum power corresponding to the \textit{null} position.

			The gradient SRP-PHAT in tandem with the MVAD form an integrated framework of multi-source localization and voice activity detection. The experiments carried out on real data recordings show that this framework is very effective in practical applications of hands-free communication.
IDIAP			
17	Free Your Hands From Prior Knowledge: A Multiclass Heterogeneous Transfer Learning Algorithm	Luo Jie, Tatiana Tommasi, Barbara Caputo	The vast majority of transfer learning methods proposed in the visual recognition domain over the last years addresses the problem of object category detection, assuming a strong control over the priors from which transfer is done. This is a strict condition, as it concretely limits the use of this type of approach in several settings: for instance, it does not allow in general to use off-the-shelf models as priors. Moreover, Moreover the lack of a multi-class formulation for most of the existing transfer learning algorithms prevents using them for object categorization problems, where their use might be beneficial especially when the number of categories grows and it becomes harder to get enough annotated data for training standard learning methods. This paper presents a multi-class transfer learning algorithm that allows to take advantage from priors built over different features and with different learning methods than what used for learning the new task. We use the priors as experts, and transfer their outputs over the new incoming samples as additional information. We cast the learning problem within the Multi Kernel Learning framework. The resulting formulation solves efficiently a joint optimization problem that determines from where and how much to transfer, with a principled multi-class formulation. Extensive experiments illustrate the value of the approach.
18	Recent Advances in Multilingual Speech Processing at Idiap Research Institute	David Imseng	We present three multilingual approaches, recently developed at Idiap Research Institute. 1) A language identification 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. 2) A theoretical framework to combine monolingual phoneme posterior probabilities in a principled way by using statistical evidence about the language identity. The framework is particularly useful to estimate multilingual phoneme posterior probabilities for mixed language speech recognition. 3) A Kullback-Leibler divergence based method that is able to exploit multilingual phoneme posterior probabilities to build speech recognition systems for resource-constrained target languages or tasks.
19	Automatic Personality Perception from Prosody	Gelareh Mohammadi, Alessandro Vinciarelli	This poster investigates an automatic approach to predict the personality of speaker from nonverbal vocal behavior namely prosodic features. 11 assessors have judged the personality of speakers in 640 audio clips in terms of Big Five model of personality. The prediction results show accuracy between 62% and 74% for different traits.
20	Codices overview	Edgar Roman-Rangel	We present an overview of the advances achieved so far in the project CODICES.

21	Anti-spoofing in 2D Face Recognition using Face-Centric Motion Correlation	Chakka Murali Mohan, Andre Anjos, Sebastien Marcel,	Face recognition has been an active research topic in the last two decades and its techniques are currently deployed in access control systems. However, spoofing attacks is a major threat causing problems to face recognition to be used as a biometrics for high-security applications. Spoofing identities using photographs, videos, masks are some of the techniques to attack 2D face recognition system. In this poster, we investigate the use of optical flow as a motion based technique to detect spoof attacks. The idea behind the motion based technique to identify spoofs is, face region moves in different direction compared to background region when a client is trying to authenticate, where as background and face regions move in similar direction in case of spoof attacks. We have used Idiap research institute's replay attacks database for our experiments. Half total error rate (HTER) on the test data set clearly shows that optical flow is a potential motion based technique to counter measure spoof attacks in 2D face recognition systems.
22	Grapheme-based Automatic Speech Recognition using KL-HMM	Ramya Rasipuram and Mathew Magimai Doss	The use of graphemes as subword units or speech recognition is interesting for reasons such as easy pronunciation generation, multilingual and crosslingual portability etc. However, modelling grapheme subword units in standard automatic speech recognition (ASR) system is not always trivial, especially for languages where the correspondence between grapheme and phoneme is weak (e.g., English). At Idiap, we are currently working on using graphemes as subword units in the framework of Kullback-Leibler divergence-based hidden Markov model (KL-HMM). More specifically, in this framework the states of the HMM represent grapheme units and the phonetic information is captured through phone posterior probabilities estimated using a multilayer perceptron serve as feature observations. In this poster, we present our investigations on (1) English language ASR, (2) multi-accent non-native speech recognition, and (3) pronunciation modelling, i.e. extraction of alternate pronunciation variants.
23	Extracting and Locating Temporal Motifs in Video Scenes Using a Hierarchical Non Parametric Bayesian Model	Rémi Emonet, Jagannadan Varadarajan, Jean-Marc Odobez	In this paper, we present an unsupervised method for mining activities in videos. From unlabeled video sequences of a scene, our method can automatically recover what are the recurrent temporal activity patterns (or motifs) and when they occur. Using non parametric Bayesian methods, we are able to automatically find both the underlying number of motifs and the number of motif occurrences in each document. The model's robustness is first validated on synthetic data. It is then applied on a large set of video data from state-of-the-art papers. We show that it can effectively recover temporal activities with high semantics for humans and strong temporal information. The model is also used for prediction where it is shown to be as efficient as other approaches. Although illustrated on video sequences, this model can be directly applied to various kinds of time series where multiple activities occur simultaneously.
24	Cancelled	Cancelled	Cancelled

25	Multi-Human Tracking and Joint Behavior Cue Estimation in Surveillance Video	Cheng Chen, Alexandre Heili, Jean-Marc Odobez	The automatic analysis and understanding of behavior and interactions is a crucial task in the design of socially intelligent video surveillance systems. Such an analysis often relies on the extraction of people behavioral cues, amongst which body pose and head pose are probably the most important ones. In this paper, we propose an approach that jointly estimates these two cues from surveillance video. Given a human track, our algorithm works in two steps. First, a per-frame analysis is conducted, in which the head is localized, head and body features are extracted, and their likelihoods under different poses is evaluated. These likelihoods are then fused within a temporal filtering framework that jointly estimate the body position, body pose and head pose by taking advantage of the soft couplings between body position (movement direction), body pose and head pose. Quantitative as well as qualitative experiments show the benefit of several aspects of our approach and in particular the benefit of the joint estimation framework for tracking the behavior cues. Further analysis of behavior and interaction could then be conducted based on the output of our system.
26	Multimodal Cue Detection Engine for Orchestrated Entertainment	Danil Korchagin, Stefan Duffner, Petr Motlicek, Carl Scheffler	In this poster, we describe a low delay real-time multimodal cue detection engine for a living room environment. The system is designed to be used in open, unconstrained environments to allow multiple people to enter, interact and leave the observable world with no constraints. It comprises detection and tracking of up to 4 faces, estimation of head poses and visual focus of attention, detection and localisation of verbal and paralinguistic events, their association and fusion. The system is designed as a coupled component for orchestrated video conferencing system to improve the overall experience of interaction between spatially separated families and friends.
27	Nonverbal Behavior of Emergent Leaders in Small Groups	Dairazalia Sanchez-Cortes	We present an analysis on how an emergent leader is perceived in newly formed small groups, and correlations between perception of leadership and automatically extracted nonverbal communicative cues. The difference in individual nonverbal features between emergent leaders and non-emergent leaders is significant and measurable using speech activity.
28	Privacy-Sensitive Audio Features for Speaker Diarization	Sree Hari Krishnan Parthasarathi	We present a comprehensive study of linear prediction residual for speaker diarization on single and multiple distant microphone conditions in privacy-sensitive settings, a requirement to analyze a wide range of spontaneous conversations. Two representations of the residual are compared, namely real-cepstrum and MFCC, with the latter performing better. Experiments on RT06eval show that residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope yields a performance close to traditional MFCC features. As a way to objectively evaluate privacy in terms of linguistic information, we perform phoneme recognition. Residual features yield low phoneme accuracies compared to traditional MFCC features.
29	GroupUs : Discovering Real-Life Interaction Types from Smartphone Proximity Data	Trinh Minh Tri Do	There is an increasing interest in analyzing social interaction from mobile sensor data, and smartphones are rapidly becoming the most attractive sensing option. We propose a new probabilistic relational model to analyze long-term dynamic social networks created by physical proximity of people. Our model can infer different interaction types from the network, revealing the participants of a given group interaction, and discovering a variety of social contexts. Our analysis is conducted on Bluetooth data sensed with smartphones for over one year on the life of 40 individuals related by professional or personal links. We objectively validate our model by studying its predictive performance, showing a significant advantage over a recently proposed model.

30	Inter-Session Variability Modelling and Joint Factor Analysis for Face Authentication	Chris McCool	We apply inter-session variability modelling and joint factor analysis to face authentication using Gaussian mixture models. These techniques, originally developed for speaker authentication, aim to explicitly model and remove detrimental within-client (inter-session) variation from client models. We apply the techniques to face authentication on the publicly-available BANCA, SCface and MOBIO databases. We propose a face authentication protocol for the challenging SCface database, and provide the first results on the MOBIO still face protocol. The techniques provide relative reductions in error rate of up to 44%, using only limited training data. On the BANCA database, our results represent a 31% reduction in error rate when benchmarked against previous work.
31	Sensing Organizational Nonverbal Behavior: the Data Collection	Laurent Nguyen and Gokul Chittaranjan	The project "Sensing Organizational Nonverbal Behavior" (SONVB) aims to address three key interrelated aspects of behavior in organizations: leadership, personality and performance. To achieve this, we plan to examine two social contexts, namely dyadic interaction in the laboratory and group interaction in real life. The laboratory part will consist of dyadic face-to-face job interviews recorded through multiple modalities (HD video, audio and depth). The real-life part on the other hand will be recorded with mobile sensors and will constitute a marketing job conducted on the street. The dataset will comprise of 50 participants. In this poster, we present our progress with this data collection effort.
32	Sparse Component Analysis for Next Generation Speech Recognition	Afsaneh Asaei, Hervé Bourlard	This research takes place in the general context of improving the performance of the Distant Speech Recognition (DSR) systems, tackling the reverberation and recognition of overlapping speech. Auditory modelling indicates that sparse representation exists in the auditory cortex. The present project thus revolves around the key question: How the sparse model could help speech recognition systems to achieve robustness in real-life applications? This study highlights two key observations: (1) Information bearing components for speech recognition are sparse in spectro-temporal domain and (2) Sparsity is preserved in reverberant acoustic conditions. Relying on these observations, we show how to cast the under-determined convolutive speech recovery as sparse approximation where we leverage recent algorithmic advances in compressive sensing and model-based sparse recovery. The results provide compelling evidence of the effectiveness of sparse recovery formulations in speech recognition.
33	Rapid adaptation for statistical speech synthesis	Lakshmi Saheer	Recent research has demonstrated the effectiveness of vocal tract length normalization (VTLN) as a rapid adaptation technique for statistical parametric speech synthesis. VTLN produces speech with naturalness preferable to that of MLLRbased adaptation techniques, being much closer in quality to that generated by the original average voice model. By contrast, with just a single parameter, VTLN captures very few speaker specific characteristics when compared to the available linear transform based adaptation techniques. This work proposes that the merits of VTLN can be combined with that of linear transform based adaptation technique in a Bayesian framework, where VTLN is used as the prior information. A novel technique of propagating the gender information from the VTLN prior through the constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation is presented. Experiments show that the resulting transformation has improved speech quality with better naturalness, intelligibility and improved speaker similarity.

34	Analyzing Big-Five Personality Traits with Large-Scale Smartphone Data	Gokul Chittaranjan	We investigate the relationship between behavioral characteristics derived from rich smartphone data and self-reported personality traits. Our data stems from smartphones of a set of 83 individuals collected over a continuous period of 8 months. From the analysis, we show that aggregated features obtained from smartphone usage data can be indicators of the Big-Five personality traits. Additionally, we develop an automatic method to infer the personality type of a user based on cellphone usage using supervised learning.
35	Efficient Boosting of multiple feature families	Charles Dubout, François Fleuret	Using multiple families of image features is a very efficient strategy to improve performance in object detection or recognition. However, such a strategy induces multiple challenges for machine learning methods, both from a computational and a statistical perspective. Our main contribution is a novel feature sampling procedure dubbed "Tasting" to improve the efficiency of Boosting in such a context. Instead of sampling features in a uniform manner, Tasting continuously estimates the expected loss reduction for each family from a limited set of features sampled prior to the learning, and biases the sampling accordingly. We evaluate the performance of this procedure with tens of families of features on four image classification and object detection data-sets. We show that Tasting, which does not require the tuning of any meta-parameter, outperforms systematically variants of uniform sampling and stat-of-the-art approaches based on bandit strategies.
36	HEAT: Iterative Relevance Feedback with One Million Images	Nicolae Suditu, François Fleuret	It has been shown repeatedly that iterative relevance feedback is a very efficient solution for content-based image retrieval. However, no existing system scales gracefully to hundreds of thousands or millions of images. We present a new approach dubbed Hierarchical and Expandable Adaptive Trace (HEAT) to tackle this problem. Our approach modulates on-the-fly the resolution of the interactive search in different parts of the image collection, by relying on a hierarchical organization of the images computed off-line. Internally, the strategy is to maintain an accurate approximation of the probabilities of relevance of the individual images while fixing an upper bound on the required computation. Our system is compared on the ImageNet database to the state-of-the-art approach it extends, by conducting user evaluations on a sub-collection of 33,000 images. Its scalability is then demonstrated by conducting similar evaluations on 1,000,000 images.
37	Posterior Features for Template-based ASR	Serena Soldo	We investigate the use of phoneme class conditional probabilities as features (posterior features) for template-based ASR. Using 75 words and 600 words task-independent and speaker-independent setup on Phonebook database, we investigate the use of different posterior distribution estimators, different distance measures that are better suited for posterior distributions, and different training data. The reported experiments clearly demonstrate that posterior features are always superior, and generalize better than other classical acoustic features (at the cost of training a posterior distribution estimator).
38	A principled approach to remove false alarms by modelling the context of a face detector	Cosmin Atanasoaei	We present a method to post-process object detections by modelling the output of an object classifier - the detection distribution around a target sub-window (context), to discriminate between false alarms and true detections. This results in a significantly reduced number of false acceptances while keeping the detection rate at approximately the same level.

39	Face Verification using Gabor Filtering and Adapted Gaussian Mixture Models	Laurent El Shafey, Roy Wallace and Sébastien Marcel	Current face recognition systems offer good performance when pose, illumination and expression conditions are controlled. However, one of the main challenges is developing systems able to work under uncontrolled conditions. We propose a new face verification scheme combining the strengths of Gabor filtering with Gaussian Mixture Model (GMM) modelling and evaluate the approach on the challenging BANCA and MOBIO databases. The proposed method reduces verification error rate by up to 52% compared to the standard GMM approach, and outperforms the state-of-the-art Local Gabor Binary Pattern Histogram Sequence (LGBPHS) technique in several scenarios.
40	Phoneme Recognition using Boosted Binary Features	Anindya Roy	In this work, we propose a novel parts-based binary-valued feature for ASR. This feature is extracted using boosted ensembles of simple threshold-based classifiers. Each such classifier looks at a specific pair of time-frequency bins located on the spectro-temporal plane. These features termed as Boosted Binary Features (BBF) are integrated into standard HMM-based system by using multilayer perceptron (MLP) and single layer perceptron (SLP). Preliminary studies on TIMIT phoneme recognition task show that BBF yields similar or better performance compared to MFCC (67.8% accuracy for BBF vs. 66.3% accuracy for MFCC) using MLP, while it yields significantly better performance than MFCC (62.8% accuracy for BBF vs. 45.9% for MFCC) using SLP. This demonstrates the potential of the proposed feature for speech recognition.
41	Recognizing the visual focus of attention in a human-robot interaction context	Samira Sheikhi	We would study the recognition of visual focus of attention in the context of human robot interaction. In our application, the robot should recognize the visual focus of attention of people while the number of them is unknown and they are free for any kind of movement and interaction.
42	OM-2: An Online Multi-class Multi-kernel Learning Algorithm	L. Jie, F. Orabona, M. Fornoni , B. Caputo, N. Cesa-Bianchi	Efficient learning from massive amounts of information is a hot topic in computer vision. Available training sets contain many examples with several visual descriptors, a setting in which current batch approaches are typically slow and does not scale well. In this work we introduce a theoretically motivated and efficient online learning algorithm for the Multi Kernel Learning (MKL) problem. For this algorithm we prove a theoretical bound on the number of multiclass mistakes made on any arbitrary data sequence. Moreover, we empirically show that its performance is on par, or better, than standard batch MKL (e.g. SILP, Sim-pleMKL) algorithms.
43	Disambiguating Temporal-Contrastive Connectives for Machine Translation	Thomas Meyer and Andrei Popescu-Belis	Temporal–contrastive discourse connectives(although, while, since, etc.) signal various types of relations between clauses such as temporal, contrast, concession and cause. They are often ambiguous and therefore difficult to translate from one language to another. We discuss several new and translation-oriented experiments for the disambiguation of a specific subset of discourse connectives in order to correct some of the translation errors made by current statistical machine translation systems.