The DIRAC AWEAR Audio-Visual Platform for Detection of Unexpected and Incongruent Events

Jörn Anemüller
University of Oldenburg
joern.anemueller@uni-oldenburg.de

Jörg-Hendrik Bach
University of Oldenburg
j.bach@uni-oldenburg.de

Barbara Caputo
IDIAP Research Institute
bcaputo@idiap.ch

Michal Havlena
Czech Technical University in Prague
havlem1@cmp.felk.cvut.cz

Luo Jie
IDIAP Research Institute
jluo@idiap.ch

Hendrik Kayser
University of Oldenburg
hendrik.kayser@uni-oldenburg.de

Bastian Leibe
ETH Zurich
leibe@vision.ee.ethz.ch

Petr Motlicek
IDIAP Research Institute
motlicek@idiap.ch

Tomas Pajdla
Czech Technical University in Prague
pajdla@cmp.felk.cvut.cz

Misha Pavel
Oregon Health & Science University
pavel@bme.ogi.edu

Akihiko Torii
Czech Technical University in Prague
torii@cmp.felk.cvut.cz

Luc Van Gool
KU Leuven and ETH Zurich
vangool@esat.kuleuven.be

Alon Zweig
Hebrew University of Jerusalem
alon.zweig@mail.huji.ac.il

Hynek Hermansky
IDIAP Research Institute
hynek@idiap.ch

ABSTRACT

It is of prime importance in everyday human life to cope with and respond appropriately to events that are not foreseen by prior experience. Machines to a large extent lack the ability to respond appropriately to such inputs. An important class of unexpected events is defined by incongruent combinations of inputs from different modalities and therefore multimodal information provides a crucial cue for the identification of such events, e.g., the sound of a voice is being heard while the person in the field-of-view does not move her lips. In the project DIRAC ("Detection and Identification of Rare Audio-visual Cues") we have been developing algorithmic approaches to the detection of such events, as well as an experimental hardware platform to test it. An audio-visual platform ("AWEAR" – audio-visual wearable device) has been constructed with the goal to help users with disabilities or a high cognitive load to deal with unexpected events. Key hardware components include stereo panoramic vision sensors and 6-channel worn-behind-the-ear (hearing aid) microphone arrays. Data have been recorded to study audio-visual tracking, a/v scene/object classification and a/v detection of incongruencies.

Categories and Subject Descriptors

H5.1 [**Multimedia Information Systems**]: Multimedia Information Systems – *Artificial, augmented, and virtual realities.*
I.5.5 [**Pattern Recognition**]: Implementation – *Interactive systems, Special architectures.*

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

General Terms: Algorithms, Design, Experimentation, Human Factors, Security.

Keywords: Augmented cognition, multimodal interaction, audio-visual, event detection, sensor platform.

1. INTRODUCTION

Under normal conditions, humans show a remarkable ability to identify and respond to unforeseen stimuli and events. Appropriate responses have frequently significant consequences (high utility), e.g., a car that suddenly approaches can lead to a potentially dangerous situation. Persons with sensory impairment (e.g., elderly) or high cognitive load (e.g., security personnel) would benefit from an assistive device that automatically detects such events and directs their attention towards them.

Algorithmic identification of unexpected events is non-trivial since they frequently do not have the properties of simple outliers. For example, individual features may be well within their normal ranges, but their combination is atypical. In this manner, the components of such stimuli may "make sense" but their combination is unexpected in certain contexts or situations. The notion of incongruencies is therefore closely linked to unexpected events, and incongruencies across modalities are particularly prominent to motivate our research of rare events detection on multimodal processing.

This contribution presents our initial progress towards developing a theoretical framework and a physical device for detection of unexpected events and highlights some results. The conceptual approach and its relevance for machine learning is outlined in section 2. We present building blocks of an audio-visual system that permits audio-visual tracking and classification in section 3. A high-level cue integration system combines audio and video streams to perform multi-modal classification and detect incongruencies across modalities (section 4). The first example applications outlined here is an audio-visual gender detection task where incongruency is to be

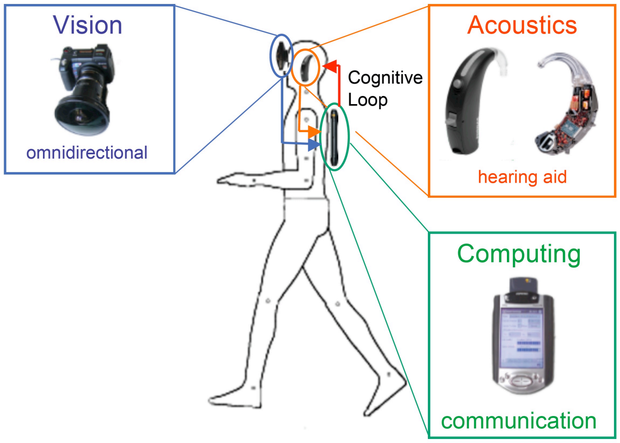

Figure 1: Schematic of the AWEAR setup.

detected when gender estimates based on visual appearance and speech characteristics diverge. Another task presented aims at audio-visual identification of in- or out-of-trusted-group subjects.

2. RARE AND INCONGRUOUS EVENTS

Machine learning systems build models of the world using training data sampled from the application domain as well as prior knowledge about the problem. These trained models are applied to new data in order to estimate the current state of the world. An implied assumption is that the future is stochastically similar to the past. This approach fails when the system is confronted with situations that are not anticipated from the past experience.

In contrast, successful natural organisms identify new, unanticipated stimuli and situations and frequently generate responses that are most appropriate in these situations. Unexpected stimuli are indicated and can be defined by incongruence between the predictions induced by the prior experience (training) and the evidence provided by the sensory data.

Our work attempts to emulate this biological ability by developing a theoretical framework for incongruent stimuli. To identify input as an incongruent stimulus, i.e., one that is not an element of a known class of objects or events, we use two parallel classifiers. The first is strongly constrained by specific knowledge (both prior and data-derived), available for a particular class of items. The second classifier is more general and less constrained, potentially comprising a superset of the objects recognizable by the more specific classifier. Both classifiers are assumed to yield class-posterior probabilities in response to a particular input signal. A sufficiently large discrepancy between posterior probabilities induced by input data in the two classifiers is taken as indication that an object or event should be considered to be incongruent.

There are various ways to incorporate prior hierarchical knowledge and constraints within different classifier levels. One approach, used to detect images of unexpected, incongruous visual objects, is to train the more general, i.e., the less constrained classifier using a larger more diverse set of stimuli, e.g., two wheeled vehicles and the other classifier using a more specific (i.e. smaller) set of more specific objects (e.g. bicycles). An incongruous item (e.g. motor bike) could then be identified by smaller posterior probability estimated by the more specific classifier relative to the probability from the more general classifier.

A different approach was applied in our work on identifying unexpected (out-of-vocabulary) lexical objects, e.g., new words [3]. The more general classifier was trained to classify (segment) speech into a sequence of phonemes, thus yielding an unconstrained sequence of phoneme labels. The more constrained classifier was trained to classify a particular set of words (highly constrained sequences of phoneme labels) from the information available in the whole spoken sentence. A word that did not belong to the expected vocabulary of the more constrained recognizer could then be identified by discrepancy in posterior probabilities of phonemes derived from both classifiers. To compare posterior probability streams, several techniques have been used, e.g. based on simple Kullback-Leiber (KL) divergence. Current version of the system is able to work with quite large vocabulary of about 5000 words.

Multimodal information streams present a related means to detect incongruous events within this framework. The unimodal classifiers are regarded as weakly constrained and their classification results are used as input for a "fusion" classifier. An incongruency between the unimodal streams will be detected as the disagreement between the more constrained fusion classifier and one of the unimodal classifiers, provided that the unimodel outputs are obtained with a sufficiently high confidence score.

3. THE AWEAR PLATFORM

In order to experiment with the proposed framework in the multimodal arena we developed the mobile audio-visual hardware platform "AWEAR" ("audio-visual wearable device", schematically depicted in Fig. 1). Extensive data recordings have been carried out in realistic environments and situations during which audio-visual data from several prototypical situations comprising audio-visual incongruent events have been obtained.

The processing pipeline of the system is shown in Fig. 2. The unimodal sensor streams are first preprocessed and then fed into detection and tracking modules. These provide the inputs for the high-level sensor fusion system that performs multimodal classification and the detection of incongruous events.

Vision data has been acquired by an omnidirectional camera consisting of the Nikon FC-E9 lens and Kyocera Finecam M410R providing 180 degrees of field of view at resolution 0.23 degrees per pixel and 3 frames per second. Omnidirectional imaging helps to monitor a large surrounding of the user in a small number of images and thus detect many events at the same time at acceptable data flow. The exotic image projection was rectified by using automatic camera calibration [8, 13] to generate perspective cutouts or cylindrical panoramas which ease further image processing and face and pedestrian detection. For moving cameras, structure from motion [9] can be used to estimate camera motion and to rectify the images as if taken by a steady camera [13].

Audio data has been recorded with a 6-channel worn-behind-the-ears microphone array that consists of two hearing aid satellites, one behind each ear and each with three microphones. The resulting system is very unobtrusive and its geometry can be considered as a hybrid incorporating bio-inspired (binaural system) and engineering elements (near-linear 3-channel sub-arrays). Data was converted to digital using an Edirol FireWire AudioCapture FA-101 AD/DA-converter. Depending on the setup of the recording situation, one or two additional channels have been recorded from close-talking lapel and headset microphones (Shure and Sennheiser, respectively).

3.1 Video Processing

On the vision side, we combine a pedestrian detector with a face detection approach in order to deliver robust performance for a range of different distances.

For pedestrian detection, we use the Implicit Shape Model (ISM) approach introduced in [10], which has been shown to work well in similar applications. This approach represents an object category by a set of local appearance features (a codebook), extracted by an interest point detector, and their learned spatial occurrence distributions. Because of the unequal camera resolution, objects that are farther away appear very small in the image, while foreground objects grow disproportionally large (and additionally suffer from distortions). Hence, several adaptations are necessary in order to apply this approach to the omnidirectional images available from the AWEAR platform. While in principle possible, it would be computationally inefficient to directly work with an omnidirectional camera geometry. Instead, we try to let the detector operate at its optimum resolution by creating a cylindrical panorama from the original omni-directional image. This way, pedestrians approaching the AWEAR setup in a 180° field of view are well visible and only show distortions when they get very close to the camera (at about 1.5m distance).

In addition, we can make several simplifying assumptions that together make detection considerably more robust. Using our knowledge about the camera setup, we can constrain pedestrian detections to lie on the ground plane. This results in a significant reduction of the search space for possible objects and thus speeds up detection. In addition, we impose a prior on plausible object sizes, which helps reduce the number of false detections.

For face detection, we use a detector based on the well-known approach by Viola & Jones [14]. The detector is applied by sliding a detection window over the image at different scales and clustering the responses. The particular detector we use is trained on frontal faces, but exhibits some tolerance to small pose changes of up to 20-30 degrees. The face detector is employed in combination with pedestrian detection. It serves two main purposes. One is to observe people at close range to the camera, where pedestrian detection may fail since only part of their body is visible. The second purpose is to identify if somebody in the surroundings is facing the AWEAR user and, if this is the case, deliver a close-up view of such a person's facial area for visual gender recognition.

3.2 Audio Processing

Preprocessing methods used for the audio stream are motivated by the fact that real audio data is characterized by its strong amplitude modulation content, i.e., signal energy exhibits a large variance when observed with a time-constant of about 30 ms. To capture the modulation structure of the sounds, signals were first decomposed into 17 different spectral "ERB" bands from about 50 Hz to about 3800 Hz with a spectral width of one ERB unit that resembles the logarithmically scaled sensitivity of human and animal auditory systems. Log-scaled signal amplitudes within each band were analyzed with a second spectral decomposition of 1 s long windows that characterized the time-scale of the amplitude modulations from 2 Hz to 30 Hz within this spectral band. Hence, the original time-domain audio signal was transformed into the 3-dimensional representation of the "amplitude modulation spectrogram" [7] with dimensions time, frequency and modulation frequency which was then employed as features for further larger margin-based classification stages for detection of sound and in particular speech sources.

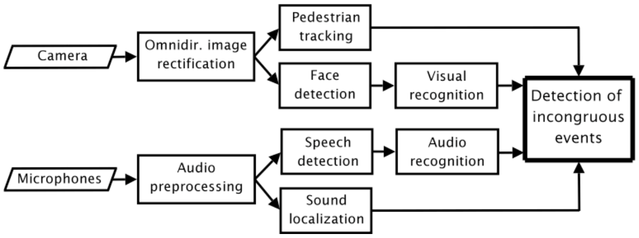

Figure 2: Processing pipeline of our system.

Tracking of audio sources is based on the DOA (direction of arrival) method that has been adapted to adequately reflect the acoustic properties of the head-worn microphone array. The basic version of the employed tracking algorithm is based on estimation of time-delays between left- and right-ear microphones and derives angular source direction estimates through the Woodworth-Schlossberg formula that compensates for the traveling time of the acoustic wave around the approximate sphere of the human head [12]. A refined version of the tracking algorithm compensates for the shading effect of the head that introduces level differences between left and right ears.

4. AUDIO-VISUAL INCONGRUENCY DETECTION

A multitude of audio-visual scenes with incongruencies across modalities has been recorded, covering in total over 100 scenes recorded with 27 speakers. One type of incongruency used pertains to localization, i.e., the spatial position or direction of a subject is different in audio and video channels. E.g., a person is appearing in the field of view at a frontal position but sound is localized as originating from the side. Another incongruency investigated is that of visual and audio appearance of gender. E.g., a male person would speak with a high-pitched voice leading to contradictory gender classification results in the different modalities.

To integrate the audio-visual inputs from the AWEAR platform for performing audio-visual tracking, a/v scene/object classification and a/v detection of incongruencies, we use the high-level integration approach. A classifier is constructed for each separate cue, each of them providing a class label estimate. All those hypotheses are then combined together to achieve a decision. In case of audio-visual tracking, the hypotheses are the predicted positions. For classification and recognition tasks, the hypotheses are confidence values for the predicted labels.

The integration strategy we applied is an extension of the weak coupling method called accumulation [5]. It is a weighted linear combination of the hypotheses on different cues. It has been shown in many cognitive and neurophysiology studies [4, 6] that humans use a similar approach for integrating multi-sensory inputs and integrate them in an optimal way. It has also been shown to achieve better performance when implemented on artificial systems [1]. The incongruent events are first defined as different classifiers giving contradicting decisions, however, both with very high confidence. To interpret these incongruencies also requires some prior-knowledge, that is, to define a proper threshold so as to minimize the false alarm due to input noise, while maintaining a high detection rate.

4.1 Detection Results

We report first results for incongruency detection on data of 30 audio-visual speaker sequences (17 speakers, 7 male and 10 female) acquired using the AWEAR platform, cf. Fig. 3 for an example snapshot. The speakers were asked to approach the camera and read

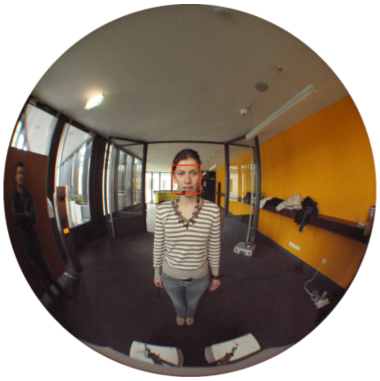

Figure 3: Example omnidirectional image for speaker verification.

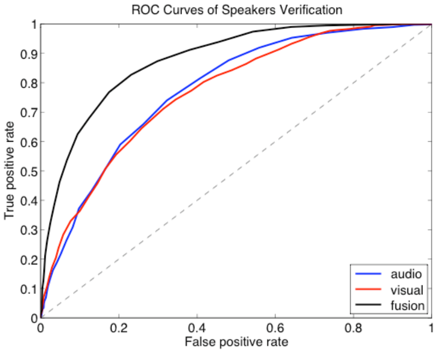

Figure 4: ROC curves for speaker verification.

a sentence about one minute long. The speech signals were captured by a head microphone worn by the actors. In a few sequences, the actors were asked to pretend an altered voice, that is, the male actors tried to speak with a high-pitched, female-like voice, and vice versa. We performed two kinds of experiments on the sequences, namely gender recognition and speaker verification, and found that integration of audio-visual cues could achieve better recognition performance than using a single modality alone, in particular under very noisy condition. For example, in some of the sequences the illumination conditions were very bad and the visual classifier gave many wrong decisions on each frame and provided low confidence in its output, while the audio classifier performed well and compensated for the weak classifier. The same effect was observed in the opposite direction when the audio chancel was noisy.

In the gender recognition task, when the speakers were using altered voices, the audio gender classifier was usually "fooled" by the voice: It's output indicated high confidence for a wrong decision, while visual gender classifier gave the opposite decision again with high confidence. In the speaker verification task, we randomly selected 6 speakers as the trusted group, and the rest of the speakers belonged to the untrusted group. Our algorithms can accurately recognize all the speakers in the trust group. In addition, an ROC curve for unknown speaker verification was obtained by varying the detection threshold (cf. Fig. 4). It shows that by integrating audio-visual cues we are able to achieve higher detection performance at lower false alarm rate.

5. CONCLUSION

The present contribution has motivated the significance of dealing with unexpected events and has proposed the use of multi-modal information to detect unexpected events that are characterized by cross-modal incongruencies. The study has been facilitated by data recorded with the AWEAR device that is intended as an audio-visual cognitive aid. First results with our biologically-inspired approach on those data indicate that multimodal information provides significant cues for continuous evaluation of the consistency of events in our environment and thereby enables humans to identify cross-modal incongruous events. Our future work will be focused on evaluating and testing the approach in more realistic situations and applications.

Acknowledgments: Support by the European Commission under integrated project DIRAC (Detection and Identification of Rare Audio-visual Cues) IST-027787 is gratefully acknowledged.

6. REFERENCES

[1] Anemüller, J., Bach, J.-H., Caputo, B., Jie, L., Ohl, F., Orabona, F., Vogels, R., Weinshall, D. and Zweig, A. 2008. Biologically motivated audio-visual cue integration for object categorization. Proc. International Conference on Cognitive Systems (CogSys 2008).

[2] Anemüller, J., Schmidt, D. and Bach, J.-H. 2008. Detection of speech embedded in real acoustic background based on amplitude modulation spectrogram features. Interspeech 2008 (to appear).

[3] Burget, L., Schwarz, P., Matejka, P., Hannemann, M., Rasrtow, A., White, C., Khudanpur, S., Hermansky, H. and Cernocky, J. Combination of strongly and weekly constrained classifiers for reliable detection of OOVs. Proc. ICASSP 2008.

[4] Burr, D. and Alais, D. 2006. Combining visual and auditory information. Progress in brain research, vol. 155.

[5] Clark, J. and Yuille, A. 1990. Data fusion for sensory information processing systems. Kluwer Academic Publisher.

[6] Ernst, M. O. and Bülthoff, H. H. 2004. Merging the senses into a robust percept. Trends in Cognitive Sciences, 8(4):162-169.

[7] Kollmeier, B. and Koch, R. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust. Soc. Am., 95(3), 1994.

[8] Kukelova, Z. and Pajdla, T. A Minimal Solution to the Autocalibration of Radial Distortion. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007.

[9] Leibe, B., Cornelis, N., Cornelis, K. and Van Gool, L. Dynamic 3D Scene Analyses from a Moving Vehicle. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007.

[10] Leibe, B., Leonardis, A., Schiele, B. Robust Object Detection with Interleaved Categorization and Segmentation. International Journal of Computer Vision, 77:259-289, 2008.

[11] Micusik, B. and Pajdla, T. Structure from motion with wide circular field of view cameras. IEEE Trans. PAMI, 28(7):1135-1149, 2006.

[12] Rohdenburg, T., Goetze, S., Hohmann, V., Kammeyer, K.-D. and Kollmeier, B. Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays. Proc. ICASSP 2008.

[13] Torii, A., Havlena, M., Pajdla, T., Leibe, B. Measuring camera translation by the dominant apical angle. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008.

[14] Viola, P. and Jones, M. Robust real-time face detection. International Journal of Computer Vision, 57(2):137–154, 2004.