Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D6.5
Detection of Audio-Visual Rare Events

Date of deliverable: 30.06.2008
Actual submission date: 20.06.2008

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **Idiap Research Institute**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

DIRAC

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D6.5 Detection of Audio-Visual Rare Events

Idiap Research Institute (IDIAP)
The Hebrew University of Jerusalem (HUJI)
Carl von Ossietzky University Oldenburg (OL)
Oregon Health & Sciences University (OHSU)

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. Here we identify distinct types of unexpected events, focusing on 'incongruent events'. 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. Another type of 'incongruent events' we consider is situations when 'general level' and 'specific level' classifiers give conflicting predictions: an event where the probability computed based on some more specific level is much smaller than the probability computed based on some more general level. 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. Data have been recorded to study audio-visual tracking, A/V scene/object classification and A/V detection of incongruencies.

Table of Content

1. Introduction

Under normal conditions, humans show a remarkable ability to identify and deal with unforeseen events. Such events are potentially associated with 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 are typically not analogous to simple outliers. Rather, they are constituted of parts that each "make sense" and only are unexpected in combination or in a certain situation. The notion of incongruencies is therefore closely linked to unexpected events, and incongruencies across modalities are particularly prominent to allow us basing rare events detection on multimodal processing.

This report present our progress towards a framework for detection of and a physical device for detection of one type of unexpected audio-visual events and highlights some results. The conceptual approach and its relevance for machine learning is outlined in Section 2. We then present building blocks of an audio-visual system that permits audiovisual 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/hierarchy (Section 4). Respectively, we show promising results with real data on two example applications: first example is an audio-visual gender detection task where incongruence is to be detected when gender estimates based on visual appearance and speech characteristics diverge, and the identification of a new sub-class (e.g., the face of a new individual) in audio-visual facial object recognition.

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. 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 inputs usually guarantee a diverse, information rich and robust inputs. There is plenty of evidence that integrating inputs from different sensory modalities can greatly enhance the ability of animals and humans to cope economically and flexibly with complex and ever-changing environment [1]. Moreover, multimodal information streams intrinsically present a related mean to detect one type of 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 incongruence 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 unimodal outputs are obtained with a sufficiently high confidence score.

3. Audio-Visual Data Processing

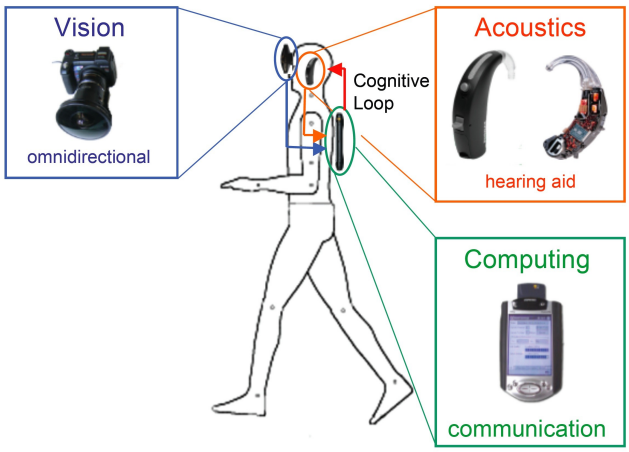

Figure 1. Schematic of the AWEAR setup

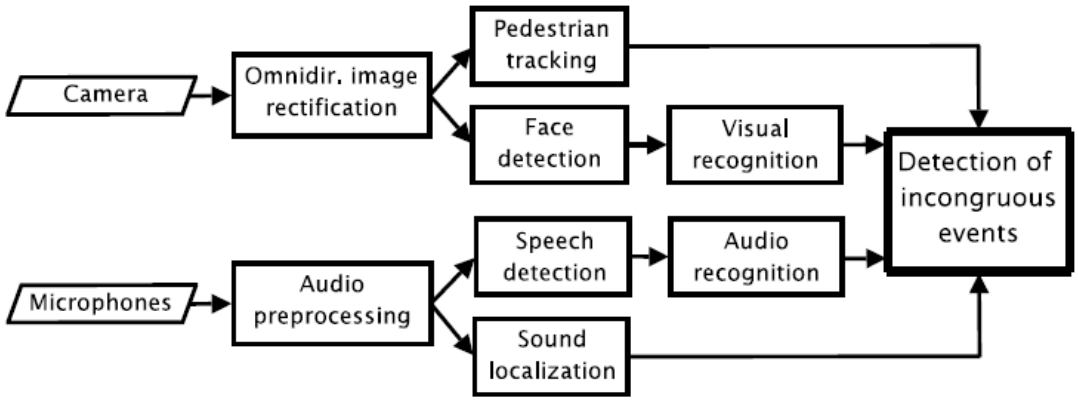

Figure 2. Processing pipeline of our system

An audio-visual platform that will ultimately be wearable ("AWEAR", schematically depicted in Fig. 1) has been constructed to study incongruous events in realistic environments and situations. Extensive data recordings have been carried out during which audio-visual data from several prototypical situations comprising audio-visual incongruent events has 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, tracking and recognition modules. These provide the inputs for the high-level sensor fusion system that performs multimodal classification and the detection of incongruous events. We will briefly describe the algorithms we used in these modules in the rest of this section.

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 Fire Wire Audio Capture 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).

Vision data has been acquired by an omni-directional 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. Omni-directional imaging helps to monitor large surrounding of the observed 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 [2, 3] to generate perspective cutouts or cylindrical panoramas which ease further image processing and face and pedestrian detection. For moving cameras, structure from motion [4] can be used to estimate camera motion and to rectify the images as if taken by steady camera [3].

4.1 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" [5] 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.

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 [6]. A refined version of the tracking algorithm compensates for the shading effect of the head that introduces level differences between left and right ears.

In the speaker verification experiments speech-non-speech detection (SND) is performed using simple logarithmic full-band energy based features employed with a large temporal context [7]. The weights of the context around the frame-to-be-classified are obtained using Linear Discriminant Analysis (LDA). This method gives us an interpretation in terms of a filter in the modulation spectral domain. And we use 26 dimensional PLP cepstral features [8] consisting of 13 static features and 13 dynamic features (first-order derivative) as speaker specific features. Though in practice it has been observed that PLP is more speaker-independent when compared to MFCC in speech recognition studies, but at the same time PLP has been observed to be more robust in noisy conditions compared to MFCC. Because of the latter reason we chose PLP features for our study.

4.2 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 [9], 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 disproportionately large (and additionally suffer from distortions). Hence, several adaptations are necessary in order to apply this approach to the omni-directional images available from the AWEAR platform. While in principle possible, it would be computationally inefficient to directly work with an omni-directional 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 and recognition, instead of using some well-known approaches for face, we use a more general object detection and recognition algorithm [10, 11], which atomically find the object regions from unsegmented images using a boosting algorithm. To learn object class classifiers we used a variant of the method in [11]. Specifically, we use an object class model as in [10] to recognize the higher more general level (face). In order to distinguish between different individual speakers, we build different one against one SVM classifiers and a majority voting method for obtain the final decision. The features of the facial images were extracted using the general category model, as described in [11]. This approach have been proved to work on various object categories. Thus, we would also like to test the same algorithm on other kinds of application.

4. Audio-Visual Rare Event Detection

4.1 Incongruence Across Modality

A multitude of audio-visual scenes with incongruencies across modalities has been recorded. One type of incongruence 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 incongruence 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. In the second example, we acquired in total 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 start a roughly one minute long conversation. The speech signals were captured by a head microphone worn by the actors, see Fig 3 for an example snapshot. 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. These sequences will serve as our main data for testing our rare event detection algorithm reported here.

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 [12]. It is a weighted linear combination of the hypotheses on different cues. It has been shown in many cognitive and neurophysiology studies [13, 14] 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 [15]. 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.

In the preliminary gender recognition experiments we performed. We 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 condition was 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: its output indicated high confidence for a wrong decision, while visual gender classifier gave the opposite decision again with high confidence.

4.2 Incongruence between General and Specific Classifier – New Face Recognition

We tested our algorithm presented in Section 2.1 on audio-visual speaker verification task. The task is to identify an individual as belonging to the trusted group of individuals vs. being unknown, i.e. known sub-class vs. new sub-class in a class membership hierarchy. In this setup, the general parent category level is the 'speech' (audio) and 'face' (visual), and the different individuals are the offspring (sub-class) levels.

We used SVM classifiers with RBF kernel as the pairwise discriminative classifiers for each of the different audio/visual representations separately. Given the classifiers trained for all levels in the hierarchies, we propose the following algorithm to decide whether an instance should be classified as known subclass, an unknown one, or neither (background/non-speech).

Test the input data on the general level classifier. If classified as a face / speech signal
a. Test the input data using all pairwise sub-class classifiers, and decide on a final classification according to a majority voting;
b. Calculate the confidence as the final classification as the average margin of all SVM classifiers contributing to the majority vote;
c. Normalize this confidence according to the average confidence computed during training phase using training samples;
d. If the confidence is lower than a certain threshold, classify as a unknown speaker. Else, give the classification results as final output.

In the experiments, we choose 6 speakers as member of the trusted group from the sequences of 17 speakers (those few sequences with altered voices were filtered out), while the rest were assumed unknown. The method was applied separately using each one of the different modalities, and also in an integrated manner using both modalities. For this fusion the audio signal and visual signal were synchronized, and the winning classification margins of both signals were normalized to the same scale and averaged to obtain a single margin for the combined method.

Since the goal is to identify novel incongruent events, true positive and false positive rates were calculated by considering all frames from the unknown test sequences as positive events and the known individual test sequences as negative events. We compared our method to novelty detection based on one-class SVM [16] extended to our multi-class case. Decision was obtained by comparing the maximal margin over all one-class classifiers to a varying threshold. As can be seen in Fig. 3, our method performs substantially better in both modalities as compared to the "standard" one class approach for novelty detection. Performance is further improved by fusing both modalities.

Figure 3. Example frame used for visual verification task

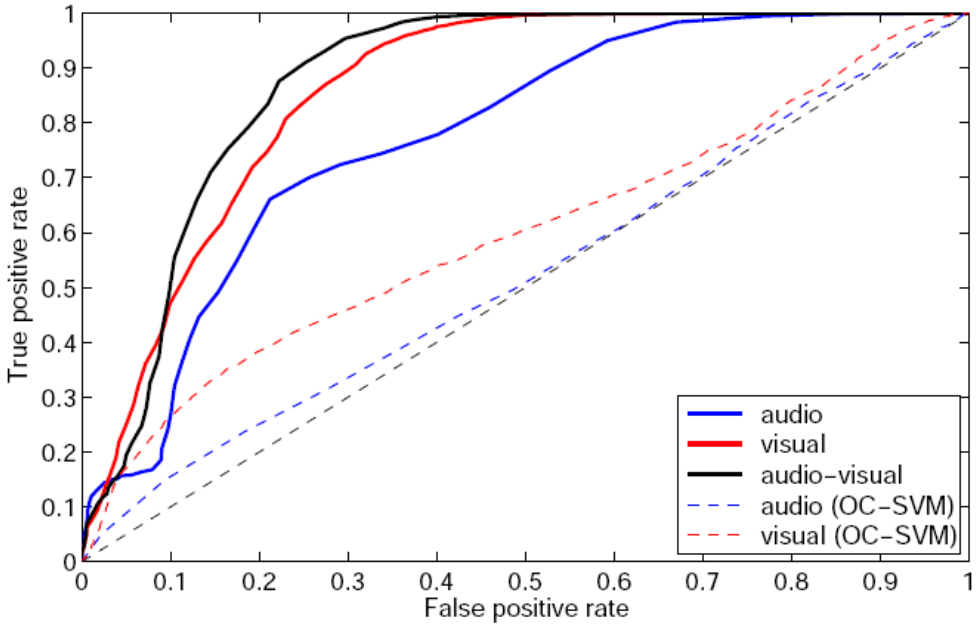

Figure 4. True Positive vs. False Positive rates when detecting unknown vs. trusted individuals. The unknown are regarded as positive events. Results are shown for the proposed method using both modalities separately and the combined method (solid lines). For comparison, we show results with a more traditional novelty detection method using One-Class SVM (dashed lines).

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 and cross hierarchical incongruencies. The study has been facilitated by data recorded with the AWEAR device that is intended as an audio-visual cognitive aid. Our first results on those data indicate that multimodal information may provide significant cues for continuous evaluation of the consistency of events in our environment and thereby enable humans to identify cross-modal incongruous events. We expect that future experiments with AWEAR recorded data will help us devise novel

algorithms that continuously build representations of the present environment, detect when something unexpected happens and alert the user of such events.

6. Reference

[1] Stein, B. E. and Meredith, M. A. The merging of the senses. MIT Press, Cambridge MA, 1993.

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

[3] 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*, 2008

[4] 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*, 2007.

[5] Kollmeier, B. and Koch, R. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *The Journal of the Acoustical Society of America* , 95(3), 1994.

[6] 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. of ICASSP 2008*, 2008

[7] Parthasarathi, S. H. K., Motlicek, P. and Hermansky, H. Exploiting contextual information for speech/non-speech detection, *Proc. of TSD 2008*, LNCS 2008, 2008

[8] Hermansky, H., Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America 87*, 1990

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

[10] Bar-Hillel, A. and Weinshall, D. Efficient learning of relational object class models. *International Journal of Computer Vision*, 2007

[11] Bar-Hillel, A.,Weinshall, D. Subordinate class recognition using relational object models. *Proc. of NIPS 19*, 2006

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

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

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

[15] Luo, J. Caputo, B. Zweig, A. Bach, J.-H. and Anemuller, J. Object Category Detection Using Audio-visual Cues. *Proceedings of the International Conference on Computer Vision System (ICVS08)*, 2008.

[16] Scholkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. *Proc. of NIPS*. Volume 12. (2000) 582–588