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.)

Project no: 027787

# DIRAC

## Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D1.7
Incongruence Detection in
Audio-Visual Processing

*Date of deliverable:2008-12-31*
*Actual submission date*: *2009-02-10*

Start date of project: 01.01.2006 Duration: 60 months

*Organization name of lead contractor for this deliverable*: ex:
**Czech Technical University CTU**

Revision 1

| Project co-funded by the European Commission within the Sixth Framework Program (2002-2006) | | |
|---|---|---|
| Dissemination Level | | |
| PU | Public | |
| PP | Restricted to other program participants (including the Commission Services) | X |
| 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.)

# INCONGRUENCE DETECTION IN AUDIO-VISUAL PROCESSING

Czech Technical University

Abstract:

Intelligent systems compare their inherent model of the universe, the "theory of the universe", with observations and measurements they make. In this work we first investigate the theory of incongruence detection developed in [27, 32] and try to see it as a mechanism for theory testing, falsification, and rectification. Next we look at some examples of incongruences in audio-visual processing, ranging form 'low-level' to 'high-level' analysis.

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.)

Table of Content

# Incongruence Detection in Audio-Visual Processing

T. Pajdla[1], L. Van Gool[2,5], M. Havlena[1],
J. Heller[1], A. Torii[1], A. Ess[2], J.-H. Bach[3], H. Kayser[3],
J. Anemüller[3], P. Van Hengel[4]

December 19, 2008

**Abstract**

Intelligent systems compare their inherent model of the universe, the "theory of the universe", with observations and measurements they make. In this work we first investigate the theory of incongruence detection developed in [27, 32] and try to see it as a mechanism for theory testing, falsification, and rectification. Next we look at some examples of incongruences in audio-visual processing, ranging from 'low-level' to 'high-level' analysis.

## 1 Overview of the report

This paper consists of two major parts. Section 2 discusses the basic ideas underlying the work. These tie in with and extend those expounded in earlier documents by the Dirac consortium about its vision on how to define what 'rare events' are. In a nutshell, these are seen to correspond to incongruences between different classifiers that analyse the same data, and where some classifiers exploit stronger expectations about the world than others.

The second part of the report exemplifies the ideas and concepts of section 2 with some examples. These include experiments on the detection of incongruence in the analysis of audio-visual signals at different levels (section 3, Dirac WPs 1, 2, 5, and 6), on the detection of image acquisition anomalies (section 4, Dirac WP1), and on the detection of unexpected events at the level of tracking humans (section 5, Dirac WPs 3 and 6).

# 2 Importance of Incongruence Detection

Intelligent systems compare their inherent model of the universe, the "theory of the universe", with observations they make. The comparison of conclusions made by reasoning about well established building blocks of the theory with direct measurements associated with the conclusions allow to falsify [28] current theory and to invoke a rectification of the theory by learning from observations or restructuring the derivation scheme of the theory. It is the disagreement – incongruence – between the theory, i.e. derived conclusions, and direct observations that allows for developing a richer and better model of the world.

In particular, we investigate the theory of incongruence detection as already propounded by the Dirac consortium in earlier publications [27, 32] and then add comments about alternative forms of incongruence that figure within the scope of the same framework, but are not Dirac's focus. Also, we raise the issues on the actions that would seem an appropriate response to the different types of incongruences.

## 2.1 Theory of Incongruence in Classification and Reasoning

The contributions [27, 32] proposed an approach to modeling incongruences between classifiers which decide about occurrence of concepts (events) via two different routes of reasoning. The first way uses a single *direct* classifier trained on complete, usually complex and compound, data to decide about the presence of an event. The alternative way decides about the event by using a *composite* classifier, which combines outputs of several (in [27, 32] direct but in general possibly also other composite) classifiers in a probabilistic (logical) way.

[27, 32] assume direct classifiers to be independent, and therefore combine probabilities by multiplication for the "part-membership hierarchy", resp. by addition for the "class-membership hierarchy". Assuming a trivial probability space with values 0 and 1, this coincides with logical AND and logical OR operations. Such reasoning hence corresponds to Boolean algebra [12]. In the sequel we will look at this simplified case. The general case can be analyzed in a similar way.

According to [27, 32], the "part-membership hierarchy" concludes that there is an event iff all parts are present. For instance, to derive a dog, there must be a head, legs, and a tail present. This would be modeled in the set-theoretical model of Boolean algebra by observing a nonempty intersection of sets corresponding to the head, the legs, and the tail. In other words,

dog is the infimum of head, legs, and tail in the partial order induced by the Boolean algebra used to reason about dogs.

Analogically, "class-membership hierarchy" considers an event to occur iff at least one of its possible instances has been observed. For instance, to derive a dog, one of the known specific dog events (Afghan, Beagle, Collie) has to be present. This would be modeled by observing a nonempty union of the sets corresponding to Afghan, Beagle, Collie. The dog concept is the supremum of the Afghan, Beagle, and Collie concepts.

Independence of the direct classifiers implies that the probability $Q_a(X)$ of an event $a$ derived by a direct classifier from observation $X$ must be the same as the probability $Q^g_a(X)$, resp. $Q^s_a(X)$ derived by a combination of the probabilities $Q_{g_j}(X)$, resp. $Q_{s_i}(X)$, of other (lower level) classifiers:

$$Q_a(X) \stackrel{!}{=} Q^g_a(X) = \prod_j Q_{g_j}(X)$$

for the "part-membership hierarchy" and

$$Q_a(X) \stackrel{!}{=} Q^s_a(X) = \sum_i Q_{s_i}(X)$$

for the "class-membership hierarchy".

Disagreement can appear between different derivations of an event. In general, there are four situations:

$$Q^g_a(X) \gg Q_a(X) \qquad Q_a(X) \gg Q^g_a(X)$$
$$Q^s_a(X) \gg Q_a(X) \qquad Q_a(X) \gg Q^s_a(X)$$

In [27, 32], an observation $X$ is called incongruent iff $Q^g_a(X) \gg Q_a(X)$ or $Q_a(X) \gg Q^s_a(X)$. In other words, [27, 32] are interested in situations when $\inf_j\{g_j\} \gg a$ and $a \gg \sup_i\{s_i\}$.

In [27, 32], classifier $Q^g_a$ is considered more general than classifier $Q_a$, since it is constructed as the infimum of elements greater than $a$. Analogically, classifier $Q_a$ is considered more general than classifier $Q^s_a$, since $Q^s_a$ is constructed as the supremum of elements lower than $a$. In this way, "more general" corresponds to "greater" in the Boolean algebra. The term *specific* is dual to the term general, i.e. we can say "more specific" instead of "less general".

According to the primary focus in the Dirac project, an observation $X$ is incongruent iff the general classifier of an event fires on $X$ and the specific classifier of the event does not fire on $X$ (see [27, 32] for a more detailed discussion of this case). Table 1 offers an extension of a table from [27, 32] by showing the relationship between $Q_a$, $Q^s_a$, and $Q^g_a$.

| | | General level | Specific level | possible reason |
|---|---|---|---|---|
| 1 | $0 \simeq Q_a^g \simeq Q_a \simeq 0$<br>$0 \simeq Q_a \simeq Q_a^s \simeq 0$ | reject | reject | noisy$X$<br>new $a$ |
| 2 | $1 \simeq Q_a^g \gg Q_a \simeq 0$<br>$1 \simeq Q_a \gg Q_a^s \simeq 0$ | accept | reject | incongruent $a$ |
| 3 | $0 \simeq Q_a^g \ll Q_a \simeq 1$<br>$0 \simeq Q_a \ll Q_a^s \simeq 1$ | reject | accept | inconsistent PO<br>models wrong/incomplete |
| 4 | $1 \simeq Q_a^g \simeq Q_a \simeq 1$<br>$1 \simeq Q_a \simeq Q_a^s \simeq 1$ | accept | accept | known $a$ |

Table 1: Interpretation of agreement/disagreement of classifiers in [27, 32].

We now go through the different cases that are lined up as rows in Table 1, also in an effort to take more fully account of the presence of noise in real data, than in the initial discussion of [27, 32].

- **Case 1.** The first row, where neither the general nor the specific classifier detects the targeted pattern, points at the event's absence (background) or, otherwise, the presence of strong noise on the observations. There is little else one can then do than to collect all such cases in the hope of learning new event types, ideally in an unsupervised way. Noise will be random enough in order not to be learned as corresponding to a novel kind of event. That would at least be the expectation.

- **Case 3.** The third row in the table – "models are wrong / incomplete" – can be interpreted as "the direct classifiers $Q_{g_j}$, which are combined in the composite classifier $Q_a^g$, do not correctly capture all cases that are captured by the direct classifier $Q_a$". The situation "inconsistent with partial order" can be interpreted as "the combination rule of the composite classifier $Q_a^g$ is not sufficient". In general, when the more general classifier (which less heavily relies on prior knowledge about the world) does not respond positively to a pattern, but the more specific classifier does, then this points at a situation where the model at the general level needs to be revised. This is a situation that may happen quite often in real systems, as knowledge about the world is typically invoked to gain robustness. The more specific classifier, which calls on stronger priors one could say, may therefore survive noisy conditions, whereas the more general classifier can no longer cope. Moreover, the specific classifier may deal with a strict subset of cases which the general classifier has to handle. Here, it is useful to think of the class-membership hierarchy discussed in [27, 32]. As an example, if one has a tracker

that detects human motion as the general module, then more specific modules could be specialised on specific types of human motion, like walking, running, hopping, etc. Whereas each of the specific modules can carve out a smaller part of parameter space, the general one has the hard job of enclosing the union of quite diverse activities within its decision boundaries. Therefore, a somewhat deviant style of walking may still be picked up by the specialised walking module, which is able to sufficiently generalise within the rather small subset of patterns, but not by the general motion one. The possible action in such cases would be to take the missed pattern as a very interesting one for retraining the weaker module. This is comparable to giving an increased weight to misclassified examples in boosting. Picking up these examples wouldn't require supervision if such Dirac pipeline is implemented. The above example is an innocent version of Case 3. Worse would be if we have trained a human motion detector/tracker and a walking detector/tracker, but walking patterns would have actually never been observed by the former. Then there is something fundamentally wrong about our world model. More to this is to follow.

- **Case 2.** As said, the initial Dirac documents [27, 32], mainly discussed the situation in the 2nd row. This is indeed the relevant case when one wants to detect events that are novel or 'surprising' to the system. The general classifier can recognizes the pattern, but the specific one does not. Taking the tracking example again, the system maybe only has specific trackers for walking and running so far, but now observes hopping. The general human motion tracker responds positively, but the specific level lacks the module to deal with this novel case. Something genuinely new about the world can be learnt by the system. Again, it could collect such cases and try to cluster similar observations to create a new, specific hopping detector/tracker. We won't discuss this case further here, as indeed it has been covered extensively by the earlier Dirac documents.

- **Case 4.** This is a case where the generic and specific levels all recognize the pattern. Normally, the system is operating in cruise mode and no action is required. Yet, even here a special case may occur. Even if the general and the specific classifiers recognize the pattern, they may disagree. This case has not been highlighted in the table, but is nevertheless interesting as well. An example situation could be when two people cross each other. Blob trackers following each one of the two people may get confused and after the crossing switch targets. Trackers

informed about normal walking would object against a sudden reversal of direction and would show a higher resiliance, sticking to their original targets. Conflicting interpretations emerge at the different levels, and at least one must be wrong. Typically, the higher robustness against noise of the specific classifier may let an arbitration scheme tend to prefer its interpretation. An example from speech is when a series of phonemes are recognised by a weaker classifier and a word by a specific classifier (more informed as it knows about the language spoken). It would stand to reason to rather believe the latter. Yet, in all such cases there is a risk that such decision is wrong. If, for instance, the speaker suddenly utters a word from another language (out-of-vocabulary condition), the specific classifier would be lead astray with high probability.

All direct and composite classifiers and their relationships captured by the partial order of the Boolean algebra used to derive composite classifiers from the direct ones can be viewed as currently the best "theory of the universe". As to this theory, there is a fundamental difference between cases 2 and 3, as we will now illustrate through the relationship between $Q_a^g$ and $Q_a$. The relationship between $Q_a$ and $Q_a^s$ is similar.
We have several direct classifiers that generate events (concepts). They are related as $g_m, g_{m-1}, \ldots, g_1 > a > s_1, \ldots, s_{n-1}, s_n$. We expect $Q_a^g \simeq Q_a \simeq Q_a^s$ and thus all other cases contradict our expectation. However, it follows from the construction of $Q_a^g$, which is obtained as the greatest lower bound (infimum) of $g_j \geq a$, that $Q_a^g \geq Q_a$. Thus $Q_a^g \gg Q_a$ contradicts our expectation but is consistent with the fact that the infimum of elements greater than $a$ (provided it exists) is also greater or equal than $a$. Such large gap can be caused by "forgetting" some parts which $a$ consists of. For instance, when detecting a dog, we should often get higher support for the conjunction of head, legs, and tail than for a complete dog since we may be ignoring many important doggish properties, e.g. that the head is above the legs and far from the tail. The gap is also an indication that we should try to look for additional concepts which could be added to lower the infimum (which was too high) and thus improve our "theory of the universe".
In contradistinction, $Q_a^g \ll Q_a$ not only contradicts our expectation, but is also contradicting the theory itself since it says that the $\inf_j\{g_j\} < a$ for $g_j \geq a$, which is impossible in a consistent theory. This means that our "theory of the universe" is wrong and must be repaired, or at the least that there are flaws in our implementation. Therefore, assuming a consistent theory in a world with perfect measurements, case 3 is impossible. This does not mean that it is not interesting in practice. As already noted before, it is as it points at measures necessary to repair our view of the world, i.e. the ac-

| | | G | S | Explanation | Action |
|---|---|---|---|---|---|
| 1 | $0 \simeq Q_a^g \simeq Q_a \simeq 0$<br>$0 \simeq Q_a \simeq Q_a^s \simeq 0$ | $\times$ | $\times$ | unknown concept<br>not explained | train new direct $Q_c$<br>set $Q_c^s = 0$ & $Q_c^g = 1$ |
| 2 | $1 \simeq Q_a^g \gg Q_a \simeq 0$<br>$1 \simeq Q_a \gg Q_a^s \simeq 0$ | $\surd$ | $\times$ | known concept<br>not explained | augment $Q_a^g, Q_a^s$<br>to best explain $Q_a$ |
| 3 | $0 \simeq Q_a^g \ll Q_a \simeq 1$<br>$0 \simeq Q_a \ll Q_a^s \simeq 1$ | $\times$ | $\surd$ | inconsistent theory<br>wrong direct class. | remove links of $Q_a^g, Q_a, Q_a^s$<br>discard solitaire $Q$'s, goto 1 |
| 4 | $1 \simeq Q_a^g \simeq Q_a \simeq 1$<br>$1 \simeq Q_a \simeq Q_a^s \simeq 1$ | $\surd$ | $\surd$ | known concept<br>explained | update all classifiers |

Table 2: The interpretation of agreement/disagreement of classifiers, which can be used to build and rebuild the "theory of the universe".

tions of the classifiers. The "theory" is challenged by this case. Repairing a complete theory is likely to be much harder than just refining it. No wonder that these cases are often completely ignored [24]. However, every practical mechanism for deriving a "theory of the universe" from observations will require dealing with case 3.

Case 2 [27, 32] boils down to detecting insufficiencies of the "theory" within the "theory". Working with case 3 means to work with the "theory" from outside [15] and to be ready for rejecting the old "theory" and building a new one. From this point of view, Table 1 can be updated as shown in Table 2.

## 2.2 Building and Rebuilding a "Theory of the Universe"

It would be appealing to use incongruences and inconsistencies between classifiers as described in Table 2 for building and rebuilding a "theory of the universe".

We say that a concept $a$ is known if there exists a corresponding direct classifier $Q_a$. We say that a concept $a$ is explained if it is derived from other elements than 0 and 1, i.e. $Q_a^s \neq 0$ and $Q_a^q \neq 1$.

The last column of Table 2 presents the actions that should be taken in order to build, maintain and rebuild the "theory" from observations:

1. Initialize: $Q_a$ by $X_1$, $Q_a^s = 0$ & $Q_a^g = 1$.
2. For each new $X_k$ evaluate $Q_a^s(X_k)$, $Q_a(X_k)$, $Q_a^g(X_k)$ for all $a$ and check cases from Table 2:

case 1 (unknown concept, not explained)

- train new direct $Q_c$
- set $Q_c^s = 0$ & $Q_c^g = 1$

case 2 (known concept, not explained)

- augment $Q_a^g, Q_a^s$ to best explain $Q_a$

case 3 (inconsistent theory)

- remove links of $Q_a^g, Q_a, Q_a^s$
- discard solitaire $Q$'s, goto case 1 with c:=a

(wrong direct classifier)

- re-train direct $Q_a$

case 4 (known concept)

- update all classifiers

The above algorithm should capture the spirit of the approach rather than all the details. In particular, it is interesting to study whether and under which conditions such approach converges and how to avoid infinite cycles. It is also not clear at the moment how to distinguish incorrect theory from a wrong direct classifier.

The detection and analysis of incongruences comes almost as a necessary condition to construct intelligent audio-visual systems that work robustly. It is generally accepted that a pure bottom-up analysis from signals to semantic interpretations is not the way to go. Powerful speech recognition hinges on knowledge about a language and its words. Robust human tracking needs to exploit knowledge about human behaviour. In Dirac parlance, specific classifiers need to give feedback to weaker ones, i.e. feedback is an important control mechanism in this area just as it is in many other disciplines. With the concept of "cognitive loops", the Dirac project endeavours to also provide feedback from higher levels than usual to low levels. If all goes well, feedback will have its intended, robustifying and stabilizing effect. Yet, erroneous feedback may also amplify errors. In general, the discussion so far looks at the results of weaker and specific classifiers as independent outputs, that then can then be compared. In practice, such classifiers will often be coupled into a system, where bi-directional influences are part of the design. Finding out about incongruences should guard the system from getting into error amplifying modes. For instance, a blob tracker with some loose notion of continuous trajectories may help a walking tracker to focus on a small area of interest. The detailed, articulated motion analysis of the latter can help the blob tracker to better predict where to go next. All goes well, until the person starts to run. If the system doesn't have the notion of running (no

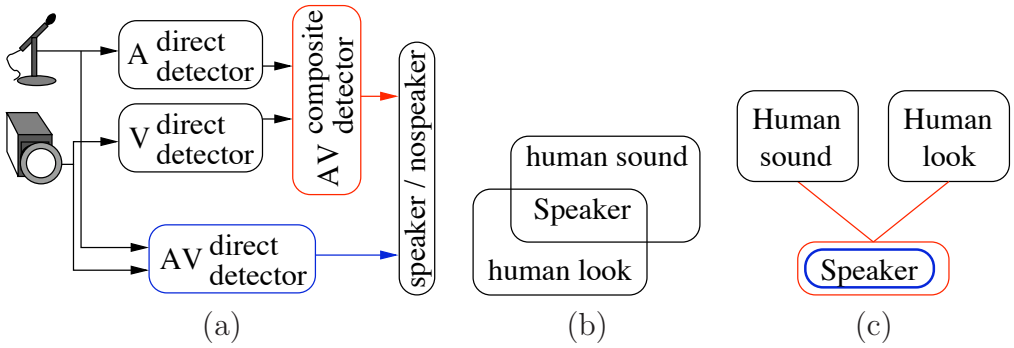



Figure 1: (a) The event "Speaker" is recognized in two ways, both by a holistic (direct) classifier, which is trained directly from complete audio-visual data, and by a composite classifier, which evaluates the conjunction of "human sound" and "human look" direct classifiers. (b) "Speaker" is given by the intersection of sets representing "human sound" and "human look", which corresponds to the infimum in the Boolean POSET (c).

specific running tracker to switch to, such switching behaviour has already been demonstrated by Dirac), then the walking tracker will hold the blob tracker back, and will actually increase the chance of the target being lost altogether. Therefore, incongruences can serve as warning flags that business-as-usual has to be aborted. Again, these are initial ideas.

To get more insights, we shall next study several examples of incongruences.

# 3 Audio-Visual Speaker Detector

As a first application of the above theory, the following section explores an audio-visual speaker detector.

## 3.1 Application of the Theory of Incongruence

Figure 1 shows an example of the "Speaker event" that is recognized in two ways, either by a direct classifier, which is trained directly from complete audio-visual data, or by a composite classifier that evaluates the conjunction of separate "Human sound event" and "Human look event" direct classifiers. "Speaker" is given by the intersection of the sets representing "human sound" and "human look" which corresponds to the infimum in the Boolean partially ordered set (POSET). In the language of [27, 32], the composite classifier

corresponds to the general level (i.e. to $Q^g_{speaker}$) while the direct classifier corresponds to the specific level (i.e. to $Q_{speaker}$).

The direct audio classifier (see Section 3.2) detects localized auditory objects, e.g. from a person speaking, in the scene. It finds the strongest response and compares it to a threshold. A side-product of the detector is the position of the strongest audio detection. The direct visual classifier (see Section 3.3) detects human body shape in an image. It finds the strongest response and compares it to a threshold. A side-product of the detector is the position of the strongest visual detection. The direct audio-visual classifier (see Section 3.4) detects the presence of a speaker. It finds the strongest response and compares it to a threshold. A side-product of the detector is the position of the strongest audio-visual detection.

The composite audio-visual classifier is constructed as the conjunction of the direct audio and visual classifiers, Figure 1. Its decision is constructed from the decisions of the separate classifiers using logical AND.

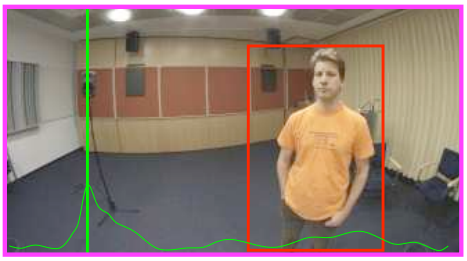
Figure 2 shows a scene with a person and a loudspeaker. Two situations are shown: (i) silent person and speaking loudspeaker, Figure 2(a), and (ii) speaking person and silent loudspeaker, Figure 2(b).

After presenting a scene with a silent person and speaking loudspeaker, the composite audio-visual classifier fires but the direct audio-visual classifier does not give a positive answer, thus creating a disagreement, incongruence, between classifiers.

This disagreement could be removed in two ways. Either the direct audio-visual classifier needs to be updated or the compound audio-visual classifier has to be modified. In the former case, a new positive example should be presented to the direct audio-visual classifier. In the latter case, a new sound-and-look concept has to be initiated. The compound audio-visual classifier is disassociated from the speaker concept, a new sound-and-look concept is created and associated to the compound classifier. This new concept will be greater than the speaker concept.

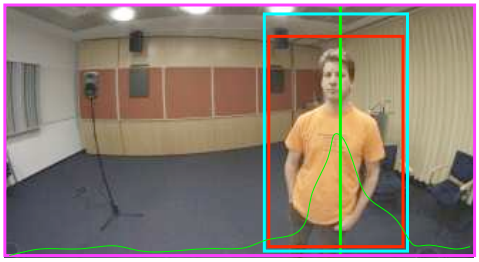
In order to derive the speaker concept from the sound-and-look concept, another concept – the concept of spatial congruence – would need to be defined, e.g., by providing its direct classifier. As the simplest direct classifier is defined by a mere list of examples and the nearest neighbor classification rule, we set positive examples to those with positive sound-and-look and with positive response of direct speaker classifier. The negative examples will then be those with positive sound-and-look response but negative direct speaker classifier response. Table 3 interprets the results of the speaker detection.

speaking loudspeaker, silent person



(a)

silent loudspeaker, speaking person



(b)

Figure 2: Conceptual sketch for speaker detection. (a) Direct audio (green) and direct visual (red) classifiers accept, so does the composite audio-visual classifier (magenta). As the direct audio-visual classifier rejects, incongruence appears. (b) If the direct audio-visual classifier (cyan) also accepts, this situation is not incongruent.

| | | Composite $Q_a^g$ (general level) | Direct $Q_a$ (specific level) | Possible reason |
|---|---|---|---|---|
| 1 | $0 \simeq Q_a^g \simeq Q_a \simeq 0$ | reject | reject | empty silent scene |
| 2 | $1 \simeq Q_a^g \gg Q_a \simeq 0$ | accept | reject | silent person speaking loudspeaker |
| 3 | $0 \simeq Q_a^g \ll Q_a \simeq 1$ | reject | accept | inconsistent POSET wrong model |
| 4 | $1 \simeq Q_a^g \simeq Q_a \simeq 1$ | accept | accept | speaking person |

Table 3: Interpretation of agreement/disagreement for example from Figure 1.

## 3.2 Direct Audio Classifier

### 3.2.1 Description of the Speech Detector

One branch of the audio-only processing pipeline is an automatic speech detector. The detector is part of a more general acoustic source detector and is based on modulation features. Specifically, we use amplitude modulation spectrograms ("AMS", [22, 11]) that have been adapted as outlined below in order to be largely invariant to speaker and channel variations. The choice of a modulation-based representation is motivated by the well-known importance in human and machine recognition of speech [8, 19] of modulation frequencies ($f_m$) in the range between 2Hz and 8Hz. Different modulation-based

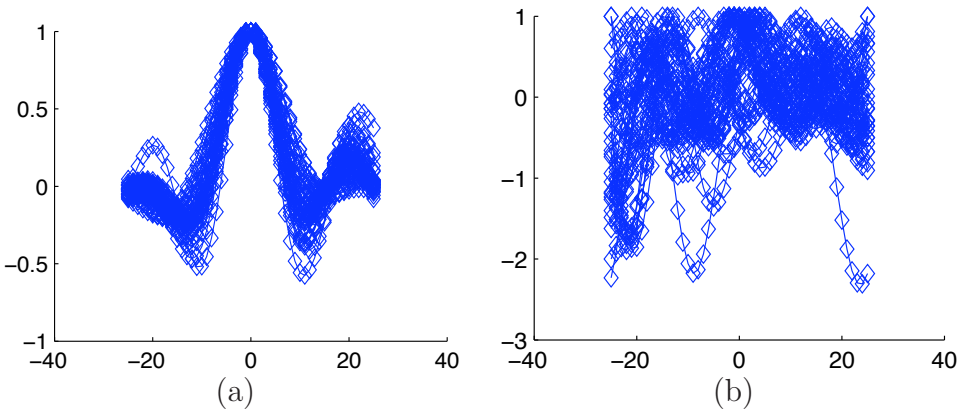



Figure 3: GCC-PHAT correlation of acoustic signal from two microphones. Signal is normalized independently in each window by dividing by its maximum. (a) Superposition of 50 positive examples. (b) Superposition of 50 negative examples.

representation have been employed for several tasks in speech processing and acoustic scene analysis, see *e.g.* [5, 26].

The AMS analyzes sound signals with respect to their modulation content by decomposing them into a 3-dimensional representation along the axes of time, frequency and modulation-frequency. The variant of the AMS employed here is tailored to be largely invariant and therefore robust with respect to (a) pitch variations between different speakers and (b) spectral distortions (channel noise) of the signal. It is computed (cf. Fig. 3) by first extracting the spectral envelopes of the acoustic signal via an FFT (32 ms Hann window, 4 ms shift), squared magnitude and Bark-band computation [33]. A subsequent logarithmic compression transforms channel noise into an additive term in each spectral band. A second FFT (1 s Hann window, 500 ms shift) is applied within each spectral band to split it into different modulation spectral bands. Additive terms are isolated into the DC-band which is subsequently discarded, thereby making the representation largely invariant with respect to channel noise. The method finishes with an envelope extraction and log-compression step. A single slice of the amplitude modulation spectrogram captures the spectral and modulation spectral information within a one second long window. By sliding the window over the signal, the temporal trajectory of modulation patterns is obtained. Our AMS decomposition uses 17 (Bark-)spectral (50 Hz to 3400 Hz) and 29 modulation-spectral (2 Hz to 30 Hz) bands, resulting in a 493-dimensional representation of the signal.

The classification backend employed consists of a standard support-vector-

machine classifier (libSVM), which was used with a linear kernel as pilot experiments did not clearly indicate a significant gain with non-linear kernels on these data. Classifier accuracy during the feature selection stage has been determined using five-fold cross-validation. The cross-validation folds have been chosen as contiguous parts of the training data to avoid artificially high cross-validation accuracy scores.

This classifier shows very good results even in very adverse conditions (SNRs down to -20dB) and outperforms a standard voice activity detection scheme by far [1].

### 3.2.2 Features for the Audio Classifier

The direct audio classifier is based on the minimization of the $L_2$ norm of the difference of the Generalized Cross Correlation with Phase Transform (GCC-PHAT) [23] and a model of a positive response. The GCC is an extension of the cross power spectral density function, which is given by the Fourier transform of the cross correlation. Given two signals $x_1(n)$ and $x_2(n)$, it is defined as:

$$G(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_1(\omega) H_2^*(\omega) \cdot X_1(\omega) X_2^*(\omega) e^{i\omega n} d\omega, \tag{1}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the respective signals and the term $H_1(\omega)H_2^*(\omega)$ denotes a general frequency weighting.
In the case of PHAT, the amplitudes of the input signals are normalized to unity, $H_1(\omega)H_2^*(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|}$:

$$G_{PHAT}(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(\omega) X_2^*(\omega)}{|X_1(\omega) X_2^*(\omega)|} e^{i\omega n} d\omega, \tag{2}$$

such that only the phase difference between the input signals is preserved. Technically, a 320-dimensional GCC-PHAT vector is provided for every video frame, covering an angular field of 180° in front of the camera having the view direction at 90°, see Figure 5.

A model of a positive GCC-PHAT response, $M$, was computed as a pointwise mean of 29 manually annotated positive responses – 51-dimensional subvectors of several GCC-PHAT vectors from training sequences, see Figure 4. Notice the peak and the ringing, the characteristic shape of a positive GCC-PHAT response.

The direct audio classifier compares the model $M$ to the scaled GCC-PHAT subvectors $G$ in a sliding window fashion:

$$E_A(x) = \sum_{k=-25}^{25} \left\| M(i) - \frac{G(x+k)}{\max_{l \in \langle -25,25 \rangle}(|G(x+l)|)} \right\|_2. \tag{3}$$

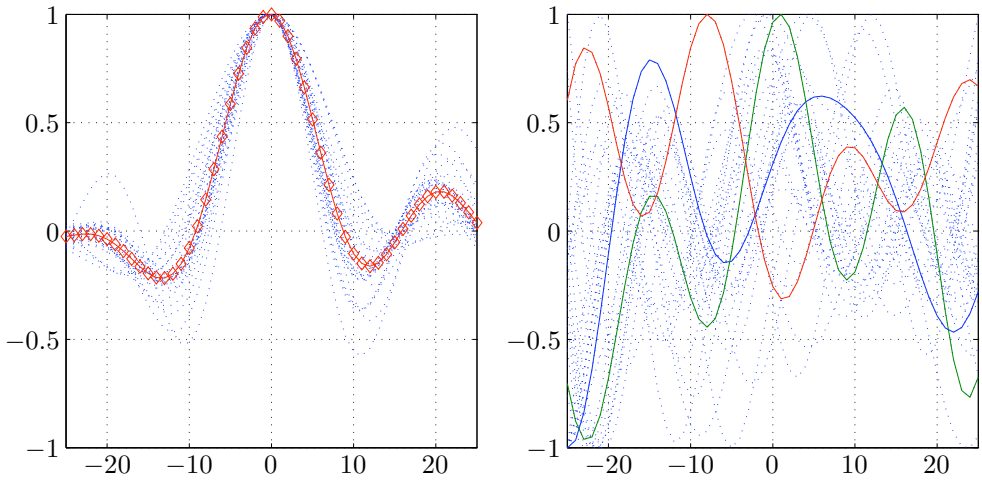



Figure 4: (a) The model of a positive GCC-PHAT response for the direct audio classifier. (b) Example of several normalized subvectors of a $G$ function classified as negative.

An example of the vector $E_A$ for a given vector GCC-PHAT $G$ is shown in Figure 5. The response of the direct audio classifier is decided by comparing $\min(E_A)$ to a threshold value $T_{E_A}$. If $G$ is a GCC-PHAT vector incident to a frame $f$:

$$\begin{aligned} \min(E_A) < T_{E_A} \quad &: \quad \text{There } \textit{is} \text{ a sound incident to the frame } f, \\ \min(E_A) \geq T_{E_A} \quad &: \quad \text{There } \textit{is no} \text{ sound incident to the frame } f. \end{aligned}$$

We trained with $T_{E_A} := 0.35$ in our experiments, which is rather conservative. Therefore several false negatives may appear.

### 3.2.3 Ongoing and Future Work

Current work concentrates on the enhancement of the GCC-algorithms and postprocessing techniques such as oversampling to increase angular resolution, short time averaging and thresholding. Preliminary experiments are promising.

With results from the GCC-based angle estimation, the implementation of an angle-dependent source predictor is the next step: The GCC is used to produce a feature vector for each angle as follows. A window is slid over the whole correlation pattern assigning the data points within the window as a feature vector to that particular angle. The feature vectors for each angle are associated with a boolean label (source present or not) to compile a labelled

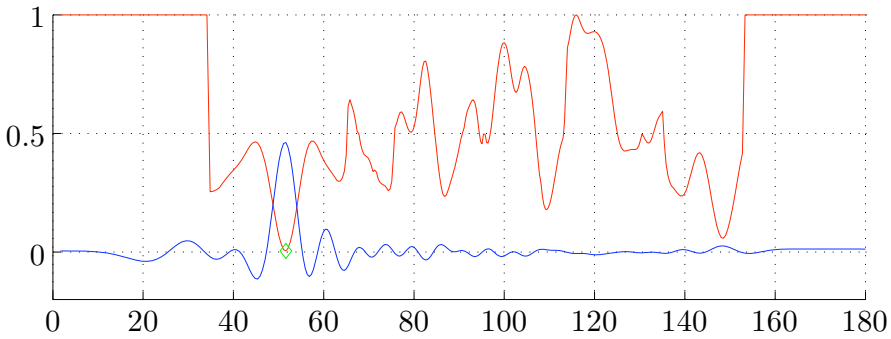



Figure 5: Blue: Example of GCC-PHAT cross-correlation of a stereo audio recording incident to a video frame. Red: $L_2$ norm of a difference of the model of a positive GCC-PHAT response and 51 degrees wide sliding window scaled to 1; green diamond represents the minimum.

data set. The resulting data set is used to train a Support Vector Machine, which will then be used to indicate appearance of sound sources at specific angles. By sliding through all available angels (-90 to +90), we construct a detector for simultaneously active localized sound sources whose number is only limited by practical concerns.

An obvious step to integrate some of the the algorithms developed so far is to couple this knowledge about source positions with the source classification framework. One conceivable approach to this end is to employ beamformer algorithms to acoustically "look" in the directions of interest as indicated by the GCC estimation. This angle-dependent data could then be fed into a slightly adopted hierarchical version of the classificator.

## 3.3 Direct Visual Classifier

Figure 8 shows a rectified large FOV image from the the AWEAR acquisition setup reprojected using cylindrical rectification [31]. The view field covers 180° horizontally.

The state-of-the-art paradigm to visual human detection [6] classifies every reasonable image window as containing a human-like shape or not, on the basis of HOG features. Here we adopt the assumption that the ground plane is parallel to the image, which can be achieved e.g. by correcting for camera motion using camera tracking and pose estimation [30]. Therefore, one can confine the search to rectangles that get taller with human proportions, like 150-190 cm tall and 50-80 cm wide, while standing on the ground plane. HoG (see Figure 6) visual features can be computed in each such rectangle as described below.

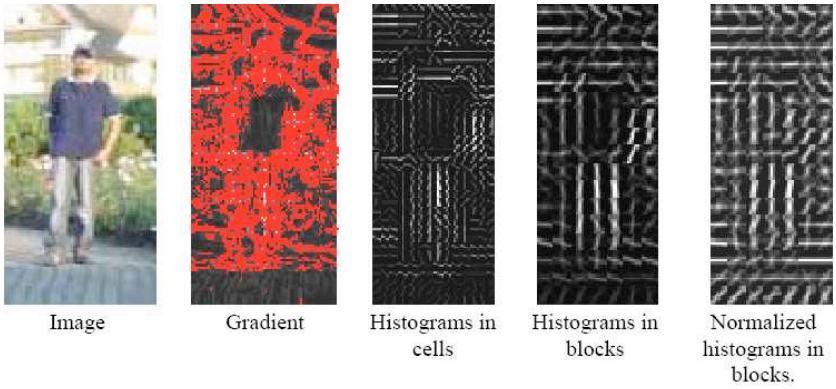



Figure 6: Histograms of oriented Gradients (HoG) features used inside the direct visual classifier. Image courtesy of [6].

The INRIA OLT detector toolkit [16] was chosen to detect the person. It is based on the histograms of oriented gradients (HoG) algorithm presented by Dalal and Triggs [6]. The method superimposes a dense grid over the detection window, producing a 3780 dimensional vector (see below). These vectors are used to train a linear SVM classifier. A detection window of $64 \times 128$ pixels is divided into cells of $8 \times 8$ pixels. Each group of $2 \times 2$ cells is then integrated into overlapping blocks. Each cell consists of 9-bin HoG that are concatenated for each block and normalized using the $L^2$ norm. Each detection window consists of $7 \times 15$ blocks resulting in 3780 features. See Figure 8 for an example of the feature vector of a detection window. As the cylindrical projection image is locally sufficiently similar to the perspective one, the detector was trained on perspective images, for which data are more readily available.

The response of the direct visual classifier is decided by comparing the confidence of the best detection to a threshold. We obtained a threshold equal to 0.01 during training, which is quite optimistic. Several false positives may appear.

## 3.4 Direct Audio-Visual Classifier

In this paragraph we describe possible ways of how to obtain a direct audio-visual classifier, i.e. a detector of objects that look like a human being producing human-like sounds, e.g. a speaking person.

Contrary to combining the results of visual detectors of humans, either generative [9] or discriminative [6], with speech sound detection [29] using Boolean logic or some uncertainty modeling, we aim at combining visual

and acoustic measurements on a lower level in a way suitable for detecting incidence of human-like shape and human-like sound. We wish to investigate inherent correlations between such shapes and sounds and test whether we can obtain a more reliable or a more specific detector compared to combining just the decisions of the two detectors. This approach has been proven useful in some other situation when detecting audio-visual events [20].

The direct combined detector takes audio and visual features computed in each angular window and returns a decision about the presence of a sounding human in that window. This can in principle be done (i) by combining features into a common feature space (normed linear space) or (ii) by combining the soft decisions (confidences) of individual classifiers together with tuning their parameters in order to achieve the best performance on training data. Next we investigate these two approaches.

### 3.4.1 Combination of Features

In principle, features can be directly combined into a common feature space in which a single (e.g. SVM) classifier can be trained. Acquiring such training data, however, is not trivial. In particular, it is difficult to record a large number of sounding people. On the other hand, it is no problem to construct such data by combining image and sound separately. Positive examples can be constructed by concatenating positive visual examples with positive audio examples, i.e. sounding human shapes. The negative examples are then constructed by the three remaining combinations, i.e. silent human shapes, as well as sounding and silent non-human shapes.

Visual and acoustic data are of a different physical nature and it thus hardly makes sense to just concatenate them into a common array. We need suitable feature transformations to obtain comparable distributions.

This approach was tested on data recorded with the AWEAR platform (see Section 3.5.1). Unfortunately, we were unable to find such transformations that would bring the features into a common normed linear space. Hence, the training of the direct audio-visual classifier using this approach was not successful.

### 3.4.2 Combination of Classifiers

In the past years, several methods to combine classifiers have been developed [21, 4, 3]. Two popular approaches worth mentioning are adaptive boosting [4] and Neyman-Pearson optimal combination [3].

These two approaches are complementary in what can be done to the classifiers that are combined. Adaptive boosting looks for the optimal linear

combination of the decisions of iteratively re-trained classifiers [4]. It is therefore sensible to use it only if the (weak) classifiers can be modified. If that is not the case, the Neyman-Pearson optimal combination is a better choice. Barruno et al. [3] presented an optimal way to combine binary classifiers in the Neyman-Pearson sense: for a given upper bound on false alarms (false positives), they find the set of combination rules maximizing the detection rate (true positives). This forms the optimal ROC curve of a combination of classifiers.

When processing AWEAR 2.0 [13] data (see Section 3.5.2), we used the following approach. Confidences of the best detections from the direct audio and direct visual detectors were put into a 4D feature vector together with their respective angular positions. The basic idea was that the classifier will infer the fact that positive training examples are those which have both confidences high and whose angular positions are equal. Since positive and negative feature vectors of this type are not linearly separable, we transformed the feature space in order to use linearly separable 3D feature vectors.

Let us examine the construction of the feature vector in greater detail. For the classifier to be able to separate two classes – "a sounding human" vs. not "a sounding human" – we construct all features so that the lower the value of the feature, the higher the confidence that the feature describes a sounding human. In the following, a feature vector $F$ for a frame $f$ will be constructed.

The first component of the feature vector represents the confidence of the best direct audio detection and is constructed from the vector $E_A$, see Equation 3. As this is actually an error measure and not a confidence, the best detection is the one with the smallest value, $\min(E_A)$.

To construct the second component of the feature vector, the best direct visual detection (i.e. the one with the highest confidence), let's denote it $c$, is selected. To follow the "the lower the value, the higher the confidence" rule, we use the reciprocal value of the confidence $\frac{1}{c}$. In case no positive detection is obtained by the direct visual detector, the second component of the feature vector is set to a random, yet sufficiently high value in order to be less confident than any true detection.

The third component of the vector $F$ exploits the proximity of the audio and visual detections. Let $\varphi \in \langle 0, 180 \rangle$ be the direction of the audio detection, i.e. $E = \min(E_A)$. Let $\psi \in \langle 0, 180 \rangle$ be the direction of the visual detection taken as the middle of the of the rectangle provided by the direct visual detector denoting the position of the human. Then the third component of the feature vector is set to $\|\varphi - \psi\|$.

To resume the previous description, the 3D feature vector $F$ for a frame

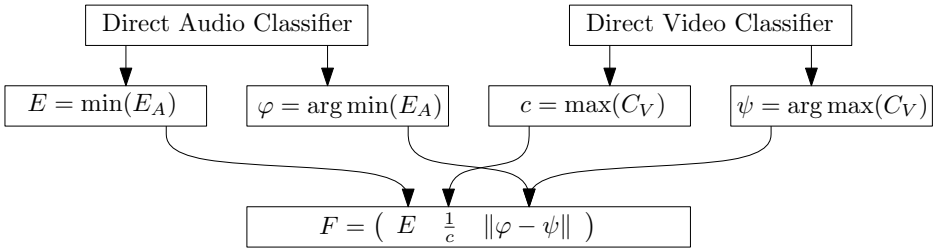



Figure 7: Construction of the 3D feature vector for the direct audio-visual classifier. First component comes from the best audio detection, second from the best visual detection, and the third one is the angular distance of the best detections.

$f$ (see Figure 7) can be schematically written as:

$$F = \left( \begin{array}{ccc} E & \frac{1}{c} & \|\varphi - \psi\| \end{array} \right).$$

## 3.5 Experimental Results

We present experiments using the direct audio-visual detector with AWEAR and AWEAR 2.0 data sequences.

### 3.5.1 AWEAR Data

The first data sets have been recorded using the AWEAR system [2]. Several 3fps sequences were acquired by two Kyocera FineCam M410R cameras with a custom mounted Nikon FC-E9, 183° FOV, lenses together with two hearing aids worn by a person standing behind the right camera. See Figure 8 for the details of the recording setup.

The GCC-PHAT was sampled with steps of 8ms, resulting in 42 samples per video frame as the cameras were set up to expose every 336ms. The 180° angle covered by the omnidirectional microphones was divided into 36 windows of 5°. The Oldenburg team also processed both channels of audio data by a speech detector resulting in a speech/no speech binary decision provided every 12ms. See Figure 9 for details on the audio timings. The audio and video streams were synchronized manually. Both video and sound data descriptions are rather idealized, since it was impossible for technical reasons to properly synchronize the sound and the video acquisition. Furthermore, the cameras did not allow for manual exposure settings, so the exposure times varied.

We experimented with the sound data in order to meaningfully combine the HoG and sound feature spaces. Since the recorded audio available for the

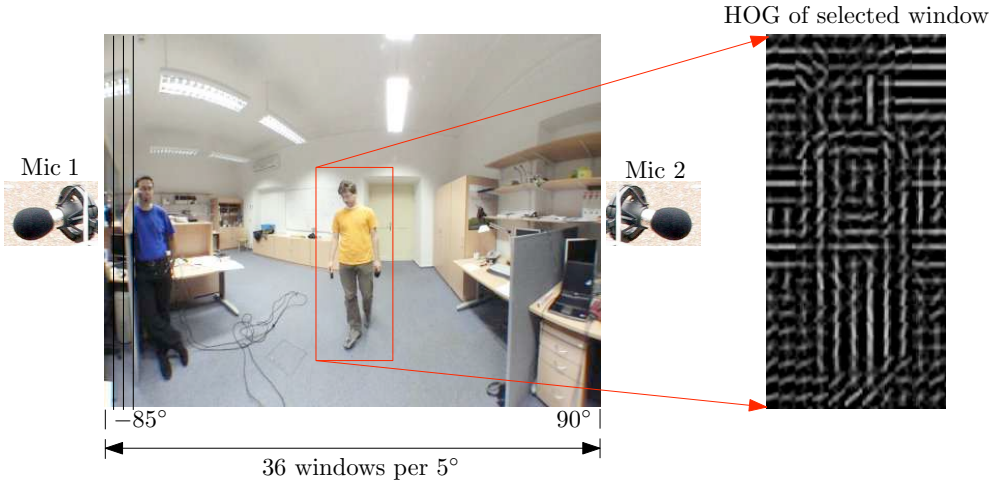



Figure 8: Left: The recording setup for AWEAR. Two Kyocera FineCam M410R cameras with custom mounted Nikon FC-E9, 183° FOV, lenses together with two hearing aids worn by a person standing behind the right camera. Right: HoG of selected window.

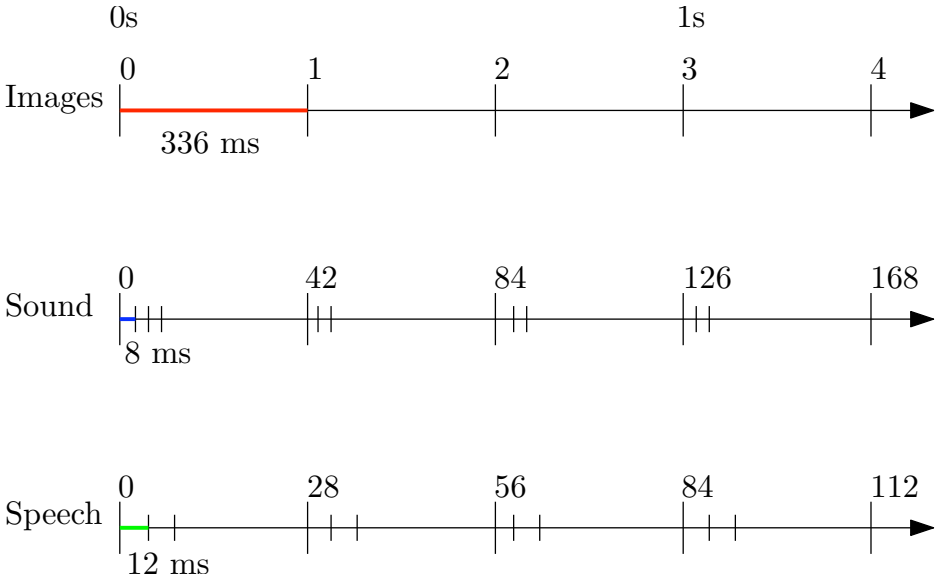



Figure 9: Audio and video timings of AWEAR sequences. As the frequency of sound data is much higher than the framerate of the cameras used, we gathered sound data incident to a video frame together.

experiments was of poor quality, we aimed to improve the discriminability of the data as well. Several combinations of averaging and thresholding were

examined. Since this preprocessing step was rather heuristic in nature, no measurements of the quality of the audio data improvement were produced. The most promising combination for the audio data preprocessing was to produce a 36-dimensional feature vector per video frame combined only from the 42 x 36 GCC-PHAT values incident to that frame's time slot. Each of the 36 values corresponded to one angular direction.

The results of the speech detector – 28 per video frame – were not considered at this stage as they were global for the whole video frame and it was impossible to obtain the angular localization of the detection as the detection results for the left and right audio channels were the same for most of the frames.

The best way to combine the audio and visual data on the feature level is yet to be researched.

### 3.5.2 AWEAR 2.0 Data

**Platform.** On November 24–26, 2008 the first recordings were made using the new AWEAR 2.0 recording platform [7]. This recording platform has been designed based on the experiences gathered with the prototype AWEAR system. The AWEAR 2.0 system, shown in Fig. 10, is more mobile and gives the ability to deal with more realistic outdoor scenes. The AWEAR 2.0 system consists of 3 computers with recording space for up to 7 hours, 2 high-resolution video cameras with fish-eye lenses, and 4 microphones connected to a high-quality audio digitizer. Due to the 4 lead-gel batteries, the autonomy is about 4 hours. The computers are networked and can be controlled wirelessly via an internet tablet. The whole system is mounted on a rigid frame backpack to ensure portability.

Specifications:

- 2 video channels  12fps effectively at 2Mpixel/image
- approximately 180 × 120∘ (HxV) field of view
- recording up to 7 audio and 1 synchronisation channels up to 96kHz sampling rate

Main components:

- 2x video recording computers with each 2GB ram, 570GB diskspace and Turion64 X2 TL-56 dual core
- 2x AVT Stingray F-201c 2Mpixel 14fps color IEEE1394b video camera
- 2x Fujinon FE185CO86HA1 fish-eye lens

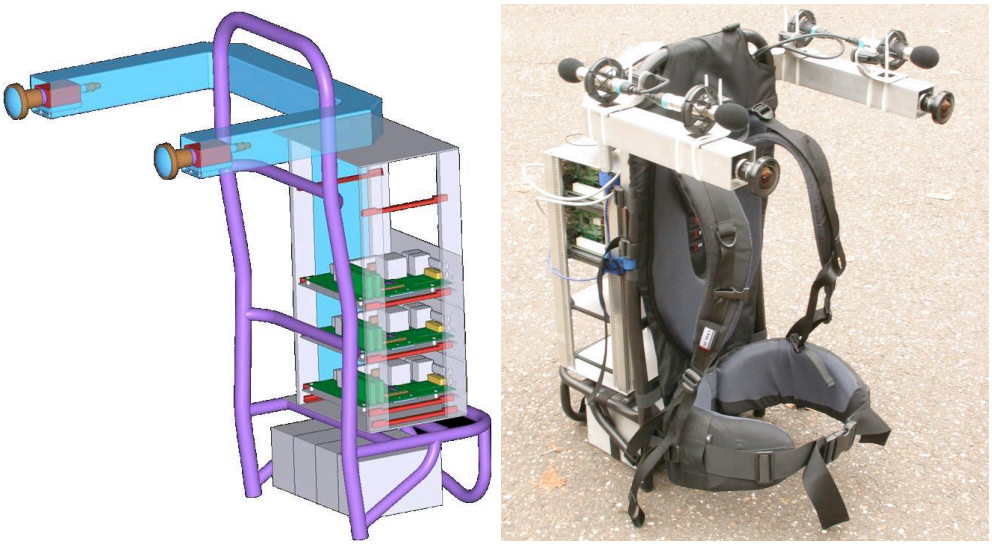


Figure 10: Left: CAD sketch of the AWEAR 2.0 platform. Right: actual platform as used for recording.

- 1x audio recording computer with 1GB ram and 250GB diskspace and Mobile Sempron 2100+
- 1x Focusrite Saffire Pro 10 I/O 8-channel 96kHz audio recording device with microphone pre-amps and phantom power supplies
- 4x T-Bone EM700 condenser microphone with directional characteristic, sensitivity -42dB (7,9 mV/Pa) and max SPL 135 dB

The first set of recordings involved one or two people in the field of view. A loudspeaker mounted on a tripod was placed on a fixed position serving as a non-human speech source. Either the person was speaking or speech was played over the loudspeaker. The first case is considered normal, whereas the second is considered incongruent, as the speech is not coming from the person. A second recording session was held on January 19, 2009. A similar setup was used as described before, cf. Fig. 11. This time the positions of loudspeaker and persons relative to the position of the AWEAR 2.0 recording system was carefully mapped, and more different persons took part.

Both recording sessions took place in the KAS room in Oldenburg, which is a specially designed room where the acoustics can be controlled by means of a multi-microphone setup, connected to a multi-channel loudspeaker system over a matrix-delay-computer steering system. In order to give optimal performance for different simulated acoustic situations, the walls, floor and

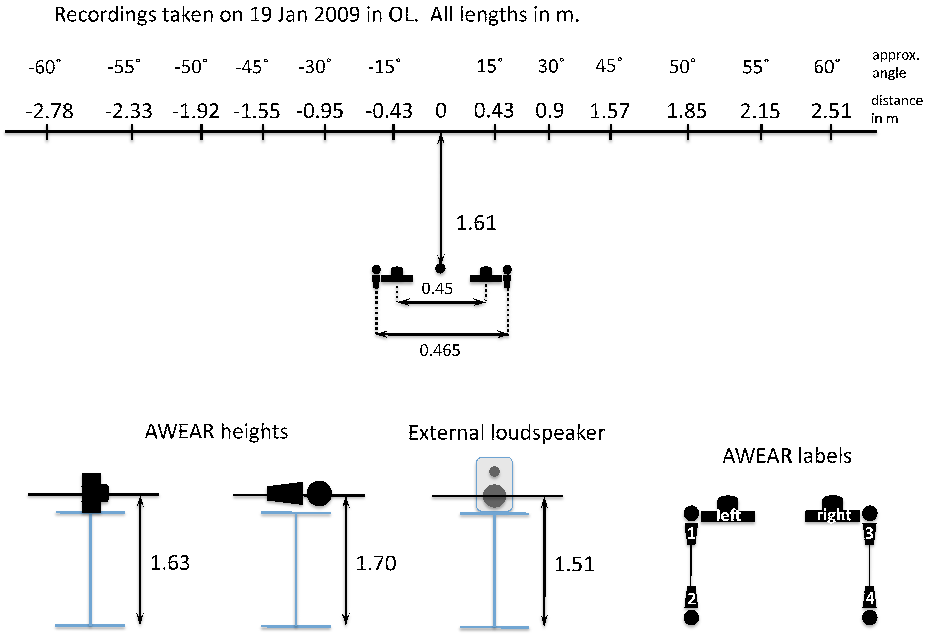



Figure 11: Layout used for Oldenburg recording on Jan 19, 2009.

ceiling of the room are made acoustically absorbing, which produces relatively little reflection in combination with relatively large size. The data and first results are available to the consortium via the DIRAC database web-page. In the recordings acquired using the AWEAR 2.0 platform, the new 4-microphone setup proved to be up to the task. The acoustic direction of arrival estimation was adapted by using a free field model (instead of the previously used HRTFs). This configuration showed improved performance due to, on the one hand, less internal noise in the microphones which provide a larger cross section surface compared to hearing aid microphones. On the other hand, the well controlled scenarios offered a constant acoustic environment yielding a better SNR throughout the recordings.

**Experiments.** First, audio-visual data consisting of a speaking person and empty silent scene were taken as the two classes and direct audio, resp. direct visual classifiers were trained to respond positively if there was a human sound, resp. a human shape detected anywhere in the image. The composite audio-visual classifier was constructed as the conjunction of the direct classifiers.

The direct audio-visual classifier is based on a combination of features provided by the INRIA OLT visual detector and the direct audio detector

| Cases from Tab. 2 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | $\neg(A\&V)\&\neg F$ | $(A\&V)\&\neg F$ | $\neg(A\&V)\&F$ | $(A\&V)\&F$ |
| SPEAKER sequence | 159 | 23 | 0 | 107 |
| LOUDSPEAKER sequence | 26 | 154 | 0 | 1 |

Table 4: Results of agreement/disagreement of classifiers for SPEAKER and LOUDSPEAKER sequences. As SPEAKER is a congruent sequence there are much more 'case 4' (speaking person) frames then 'case 2' (incongruent) frames. In LOUDSPEAKER sequence, there is only one 'case 4' frame at the very end of the sequence when the person comes to the loudspeaker but a lot of 'case 2' frames. Notice that there are no 'case 3' (wrong model) frames in both sequences.

into one 3D feature vector per video frame. Such feature vectors, annotated from several sequences captured by AWEAR 2.0 were used to train a linear classifier using $SVM^{light}$ [18]. Such a direct audio-visual classifier is able to detect up to one sounding human in an audio-visual frame. The video frames were rectified using the cylindrical projection in order to easily match the spatial positions of audio and visual detections in an angular measure, e.g. degrees.

As the video data come from the left camera, there is a discrepancy between the camera position and the apparent position of a virtual listener to which the GCC-PHAT is computed, which is the center of the acquisition platform. To compensate for this error, the distance to the sound source and the distance between the virtual listener and the camera need to be known. The listener–camera distance can be computed as 22.5cm from the known setup of the rig. The distance to the sound source is now assumed to be 1.5m from the camera. The corrected angle can then be trivially computed from the camera–listener–sound source triangle using a line–circle intersection.

The direct audio classifier, the direct visual classifier and the direct audio-visual classifier were used to classify two indoor sequences captured by the AWEAR 2.0 device. Sequence SPEAKER is a 289 frames (approximately 24s) long sequence showing a talking person, standing at the right side of the image and a silent loudspeaker at the left side. Sequence LOUDSPEAKER is a 181 frames (approximately 15s) long sequence showing a silent person standing at the right side of the image and a sounding loudspeaker at the left side. At the end of the sequence, the person moves in front of the loudspeaker. See Figure 12 and Figure 13 for sample frames from sequences SPEAKER and LOUDSPEAKER respectively.

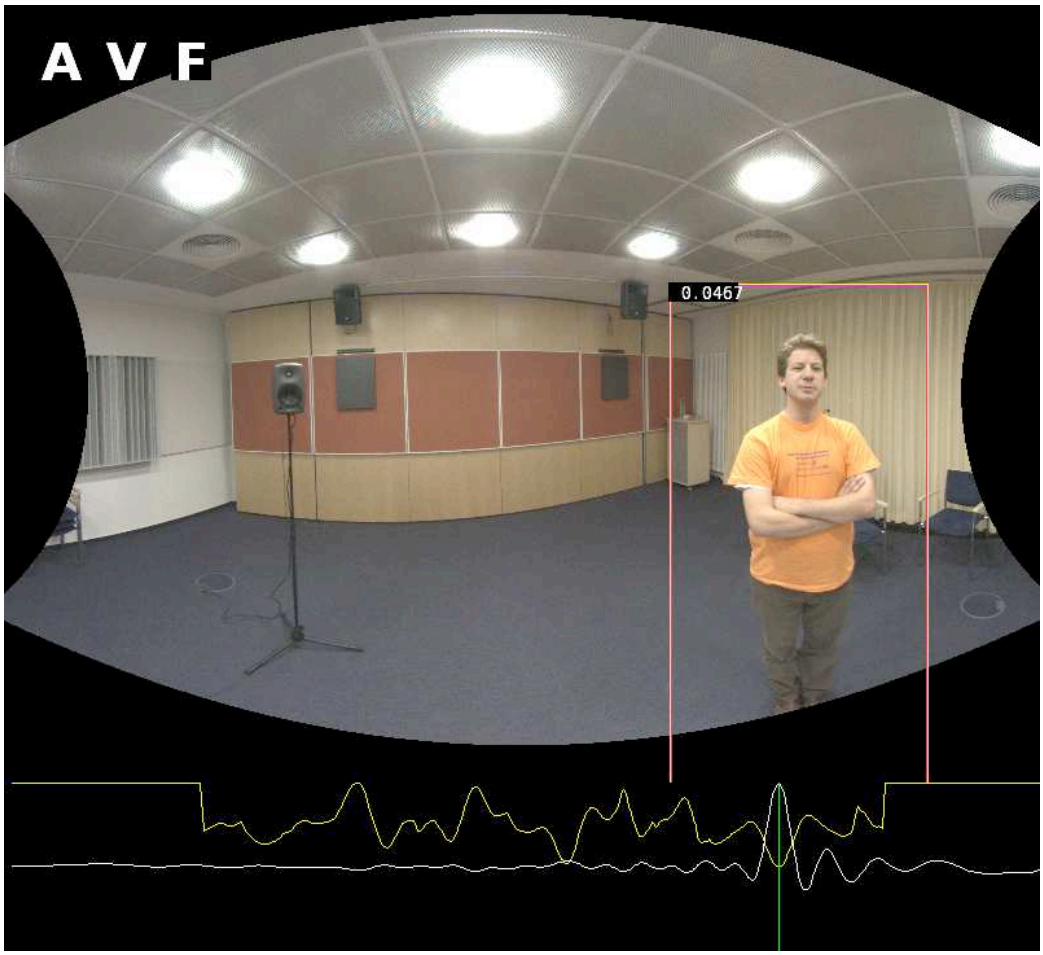



Figure 12: Sample frame from sequence SPEAKER together with the results of the classifiers. The letter 'A' denotes positive response of the direct audio classifier, the character 'V' denotes positive response of the direct visual classifier and character 'F' denotes positive response of the direct audio-visual classifier – the frame depicts a sounding human. The video frame is resampled using the cylindrical projection. The white graph represents the incident GCC-PHAT vector, see Equation 2, the yellow graph is the incident $E_A$ vector, see Equation 3, and the green line denotes the position of $\min(E_A)$. The rectangle around the person represents the position of the positive visual response.

Table 4 resumes the quantitative results of the experiment. As SPEAKER is a congruent sequence there are much more case 4 (speaking person) frames than case 2 (incongruent) frames. In the LOUDSPEAKER sequence, there is only one case 4 frame at the very end of the sequence when the person

comes to the loudspeaker but a lot of case 2 frames. Notice that there are no case 3 (wrong model) frames in either sequence.

The results could be further improved by incorporating the temporal information present in the video sequence. As the framerate is quite high, people do not move much between consecutive frames. Sound data have a temporal dependence also, furthermore, using temporal information would help to bridge negative detector responses present when the speaker breathes and therefore stops speaking for a very short while. The classical solution using a Hidden Markov Model Chain (HMMC) would assign consistent labels to individual frames if trained correctly. Using HMMC for the detectors is part of our future work, as a refinement to our incongruence detection.

# 4 Detecting Incongruences on Sensor and Low-Level Signal Processing

One of the goals of low-level processing is to preprocess incoming signals and extract reliable information. In particular, the processing should be able to detect when some of the incoming information is wrong or when it is too unreliable to be used in further processing. Why should the result of processing be wrong? There may be various reasons. For instance, one (or more) of the following problems may present themselves:

1. one or both cameras may fail to provide images,
2. one or both lenses may be out of focus,
3. one or both cameras may lose their calibration,
4. cameras may get be out of sync,
5. the epipolar geometry (EG) of the cameras may change.

The above items represent a number of events that can be detected by comparing the results of processing with expectations learned from previous situations.

Another possibility of detecting incongruences comes from the fact that AWEAR 2.0 cameras observe the same scene, only from slightly different viewpoints. Thus, the results on the left and right images should, at least for certain types of processing, be comparable.

As the processing follows the standard path, we may add a number of detectors based on statistics of the results:

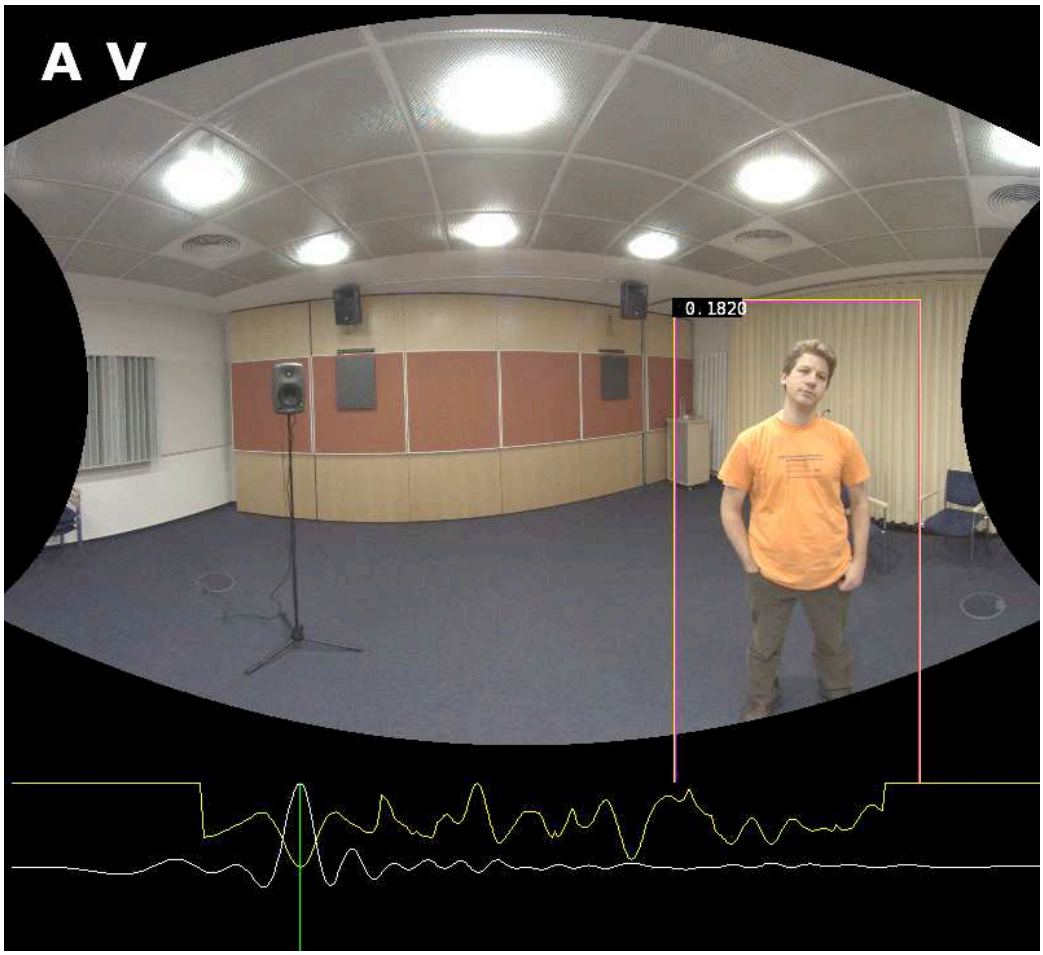



Figure 13: Sample frame from sequence LOUDSPEAKER together with the results of the classifiers. See Figure 12 for the description of the drawings and their colours. Notice the positive responses of both the direct audio and the direct visual classifiers, denoted by letters 'A' and 'V'. There is no positive response of the direct audio-visual classifier, the situation is incongruent.

1-1 Detect features in camera 1

1-2 Feature statistics in camera 1

2-1 Detect features in camera 2

2-2 Feature statistics in camera 2

12-3 Tentative matches between camera 1 & 2

12-4 Tentative matches statistics

12-5 Matches (M) by unconstrained epipolar geometry (UEG) between camera 1 & 2

12-6 Statistics on M by UEG

12-7 Matches by known EG (KEG) between cameras

12-8 Statistics on M by KEG

1-9 Tentative matches between consecutive images (CI) in camera 1

1-10 Tentative matches statistics of CI camera 1

2-9 Tentative matches between consecutive images in camera 2

2-10 Tentative matches statistics of CI camera 2

1-11 Matches by UEG in consecutive images in camera 1

1-12 Statistics on M by UEG in CI camera 1

2-11 Matches by unknown EG in consecutive images in camera 2

2-12 Statistics on M by UEG in CI camera 2

The simplest statistics (detectors, classifiers) can be constructed by looking at the number of detected features, tentative matches, and matches verified by epipolar geometries. More advanced might be various quality measures, for instance the measure based on apical angles and view field coverage as is used in randomized structure from motion (SfM) [13].

We will first design detectors of the above five problems using the number of matches. These numbers we can plot into graphs, as a function of the frame number. Since our current sequences do not exhibit abnormal events at this sensor level, we simulate the following 5 abnormal events:

1. Camera fails: replace image in camera 1 by almost black image with a small random noise.
2. Camera out of focus: blur one of the images.
3. Cameras out of sync: shift the camera 1 image stream by 3 frames w.r.t. the camera 2 image stream.
4. Camera calibration wrong: use slightly wrong camera 1 calibration.
5. Camera rig calibration wrong: use slightly wrong epipolar geometry of the rig.

Figure 14: Frames 1, 50, 100, 150, and 200 of the PEDCROSS sequence. Top row: camera 1 (left camera). Bottom row: camera 2 (right camera).

Once events (problems) are detected, the next step is to take an action that will remove (if possible) the cause of the abnormality. For instance, we can try to recalibrate individual cameras as well as the camera rig. In fact, this action might be a part of the detection as well. For instance, when we detect that we can successfully track individual cameras but cannot track the rig, we can either expect a problem in rig calibration or in synchronization. Then, we may try to recalibrate or shift frames and choose the action that will better fit to incoming data. However, if none of these actions improves the results, we can conclude that we are not dealing with these particular problems, start collecting these images, and from these try to learn a model of the situation (e.g. by clustering in the feature space provided by the statistics), which is to be added as a new phenomenon. Next we will study how much can be done by looking at the number of features and matches.

## 4.1 Detecting Abnormal Situations in Image Matching and Camera Tracking

Figure 15 shows the numbers of detected regions, tentative matches, and matches supported by epipolar geometry verification as the statistics of the PEDCROSS sequence, see Figure 14, which is 228 frames long. Figure 15(a), (b), and (c) show the statistics of the original sequence acquired under correct (normal) behavior of the system. (a) shows the number of detected SURF features. Red and blue colours correspond to camera 1 and camera 2, respectively. (b) shows the number of matches between camera 1 and camera 2 in the stereo pair, i.e. between camera 1 and 2 of the same frame. The graph shows tentative matches (red), matches supported by an unconstrained epipolar geometry (UEG) computed in every frame (blue), and a known EG

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| C1T1#Features | H | L | L | H | H | H |
| C1C1#Tent Mtch | H | L | L | H | H | H |
| C1C1#Veri Mtch UEG | H | L | L | H | H | H |
| C2T1#Features | H | H | H | H | H | H |
| C2C2#Tent Mtch | H | H | H | H | H | H |
| C2C2#Veri Mtch UEG | H | H | H | H | H | H |
| C1C2#Tent Mtch | H | L | L | H | H | H |
| C1C2#Veri Mtch UEG | H | L | L | H | H | H |
| C1C2#Veri Mtch KEG | H | L | L | L | L | L |

Table 5: Distinguishable abnormalities.

(KEG) computed using the precalibrated stereo rig configuration (green). (c) shows the number of matches between consecutive images of camera 1 and camera 2 sequences. The graph shows tentative matches of camera 1 (red) and camera 2 (blue) and matches supported by an UEG of the consecutive images of the camera 1 (magenta) and camera 2 (cyan), respectively.

The abnormal situations were simulated by tampering with the images or system parameters. In all situations, four subsequences comprising frames 51-60, 101-110, 151-160, and 201-210, resp. were modified to simulate abnormal situations. The rest of the sequence was left unaltered to visualize the impact of the abnormal situations.

The following abnormal situations were generated:

1. C1 fails: Images were 78% darkened and 1000 salt&pepper noise peaks were added, see Figure 17(b) and Figure 15(d), (e), and (f).

2. C1 out of focus : Images were artificially blurred by a 2D Gaussian filter, see Figure 17(c) and Figure 15(g), (h), and (i).

3. C1 calibration wrong : The optical center in the mathematical camera model was shifted by 10 pixels in both $x$ and $y$ direction, Figure 15(m), (n), and (o).

4. C1 C2 out of sync: Frames were desynchronized by shifting frames of one camera by three, e.g. frame 51 in camera 1 corresponded to frame 54 in camera 2, Figure 15(j), (k), and (l).

5. C1 C2 rig calibration wrong: 10deg rotation of camera 1 was applied to the precomputed stereo geometry, Figure 15(p), (q), and (r).

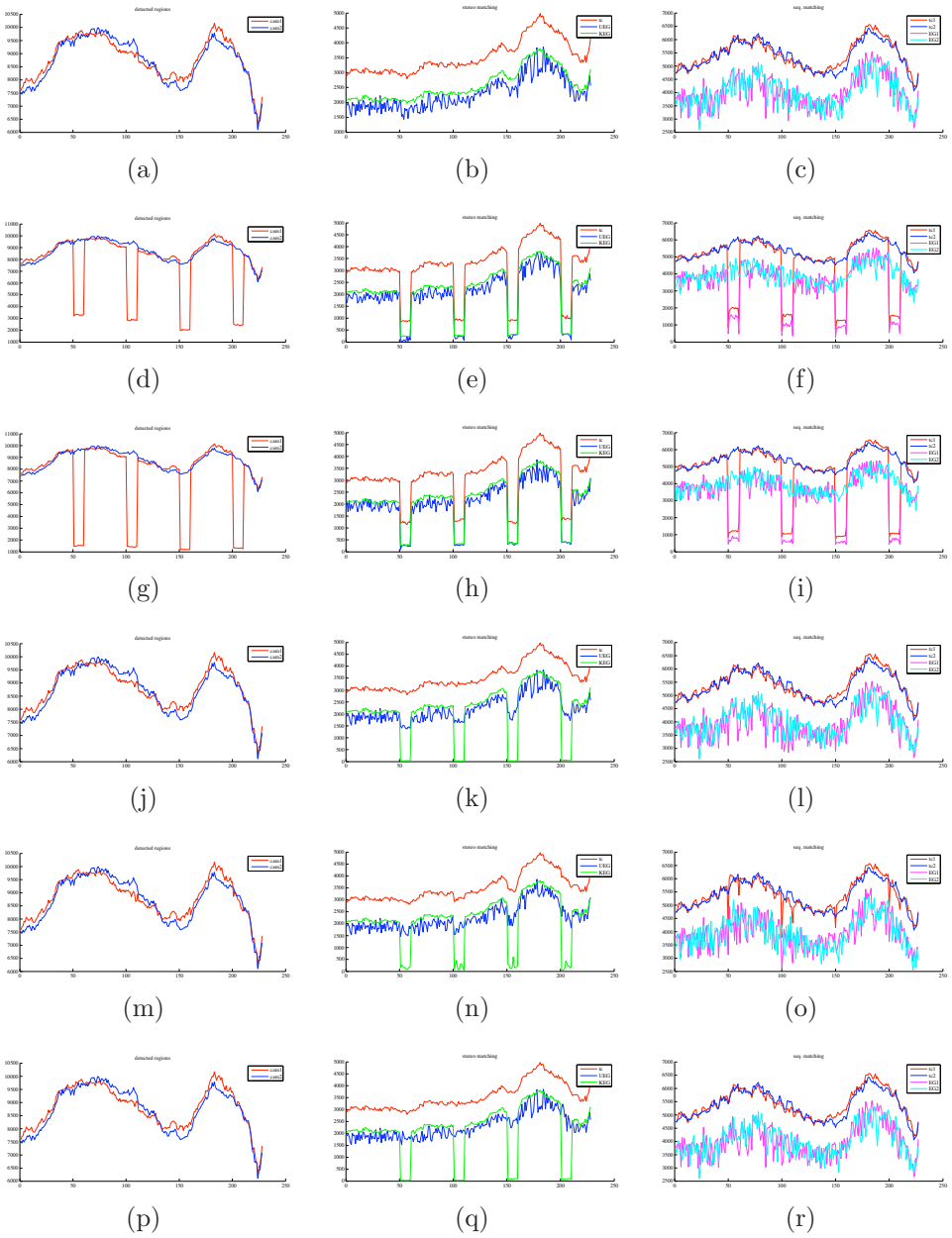


Figure 15: Abnormalities in the PEDCROSS sequence. (a) shows the number of detected SURF features. (b) shows the number of matches between camera 1 and camera 2 in the stereo pair, i.e. between camera 1 and 2 of the same frame. (c) shows the number of matches between consecutive images of camera 1 and camera 2 sequences. Colour codes are explained in the text.

The numbers of matches reported in Figure 15 yield patterns as shown in Table 5, that distinguish between 3 types of situations (corresponding to the 3 different patterns), depending on the following abnormalities:

1. correct function of the rig tracking,
2. camera 1 fails,
3. camera 1 out of focus,
4. cameras out of sync,
5. camera 1 looses the calibration,
6. wrong calibration of the camera rig,

with Ci-Ck standing for cameras i and j, Ci-Tx for camera i at time x, H for a high number, and L for a low number. The three pattern types correspond to the three sets $\{1\}$, $\{2, 3\}$, and $\{4, 5, 6\}$.

The above table can only be used when we can decide what H and L are supposed to mean. This varies from scene to scene. Therefore, we plan to develop a feature mapping from raw numbers of detections and matches into a space which is (more) invariant to scene variations.

Let us suggest to use the following features, with #F the number of extracted features, #TM the number of matches found between two images, #VMU the number of matches verified by unconstrained EG, and #VMK the number of matches verified by known EG:

1. #F(C1T1)
2. #F(C1T2)
3. #TM(C1T1,C1T2)/min(#F(C1T1),#F(C1T2))
4. #VMU(C1T1,C1T2)/#TM(C1T1,C1T2)
5. #F(C2T1)
6. #F(C2T2)
7. #TM(C2T1,C2T2)/min(#F(C2T1),#F(C2T2))
8. #VMU(C2T1,C2T2)/#TM(C22)
9. #TM(C1T1,C2T1)/min(#F(C1T1),#F(C2,T1))

10. #VMU(C1T1,C2T1)/#TM(C1T1,C2T1)

11. #VMK(C1T1,C2T1)/#TM(C1T1,C2T1)

and then train a classifier for each feature that would be able to distinguish what is L and what H. The outputs of these classifiers might be combined, e.g. by boosting.

Alternatively, we could train two SVM classifiers, one for the feature space (#F(C1T1), #F(C1T2), #F(C2T1), #F(C2T2)) and the other for the feature space of all other features in the above list. This is meaningful since the first feature space would combine absolute numbers and the second one would combine relative fractions. The output from the two classifiers would be combined by boosting.

The next open question is how to distinguish abnormalities $\{2, 3\}$ and $\{4, 5, 6\}$. There may be many ways how to learn about problems 2 and 3 or one can also try to lift problem 3 by evaluating the image sharpness and applying the appropriate inverse filter. However, these are problems that require physical interaction with the device for their true solution, and thus it should be sufficient to set an alarm and use the remaining, correct camera when it is sufficient.

Distinguishing between $\{4, 5, 6\}$ is more interesting since calibrations can be automatically updated from matches. One could suggest repairing individual camera calibrations as well as the camera rig calibration by a few steps of bundle adjustment [14] or other autocalibration techniques and compare the obtained results.

The next level of analysis of anomalies would look at the time dependences of the detections. Decisions of individual classifiers could be modeled by a Markovian Chain.

Figure 16 shows the normalization of the raw statistics from Figure 15. The normalized statistics of the original sequence is shown in (a) for the stereo matching:

(red) #TM(C1T1,C2T1)/min(#F(C1T1),#F(C2T1))

(blue) #VMU(C1T1,C2T1)/#TM(C1T1,C2T1)

(green) #VMK(C1T1,C2T1)/#TM(C1T1,C2T1)

and in (b) for the sequential matching:

(red) #TM(C1T1,C1T2)/min(#F(C1T1),#F(C1T2))

(blue) #TM(C2T1,C2T2)/min(#F(C2T1),#F(C2T2))

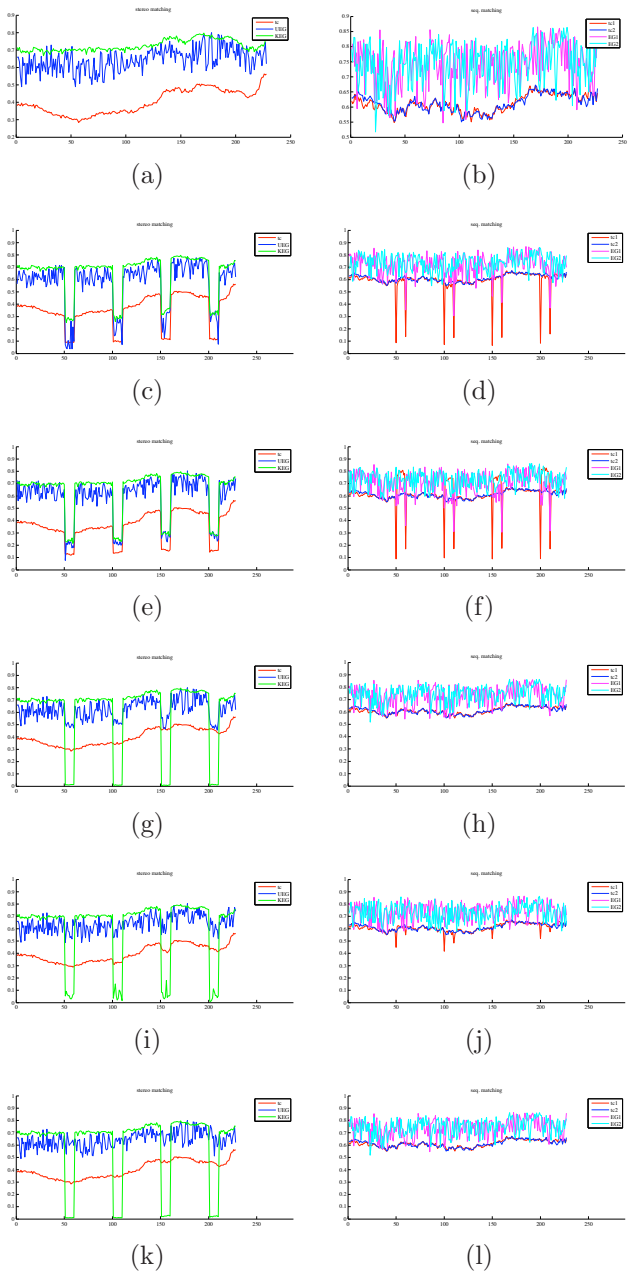


Figure 16: Normalization of the raw statistics from Figure 15. (a) for the stereo matching. (b) for the sequential matching. Colours are explained in the text.

(magenta) #VMU(C1T1,C1T2)/#TM(C1T1,C1T2)

(cyan) #VMU(C2T1,C2T2)/#TM(C2T1,C2T2)

In the same way, the second to the fifth rows correspond to the sequences of 'C1 fails', 'C1 out of focus', 'C1 calibration wrong', 'C1 C2 out of sync', and 'C1 C2 rig calibration wrong'.

We can see that the normalization rendered the values of the features more independent for the actual pair of images, i.e. for the actual scene used to compute the camera motion. We can also see that it works as a change detector in Figure 16(d) and (f). We conclude that the proposed feature transformation gave promising results.

## 4.2 Incongruences in Camera Tracking

Camera tracking provides an interesting and (in signal processing) quite common situation for detecting incongruences. The specialty of the situation lies in the fact that the results of the processing are passed from lower to higher levels. It is thus impossible, e.g., to encounter the situation when there would be many matches verified by an epipolar geometry but few tentative matches or even few detected features.

Therefore, a falsification of predicate $\#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1}) \Rightarrow \#\mathrm{F}(\mathrm{C1T1}) \wedge \#\mathrm{F}(\mathrm{C2T1})$, which would lead to the case 3, i.e. to breaking the "theory", can never be performed as it is equivalent to $\neg\#\mathrm{F}(\mathrm{C1T1}) \vee \neg\#\mathrm{F}(\mathrm{C2T1}) \Rightarrow \neg\#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1})$ being falsified by the case when there is a lack of features in one of the images but it was still possible to find a lot of tentative matches between them, which is impossible.

On the other hand, predicate $\#\mathrm{F}(\mathrm{C1T1}) \wedge \#\mathrm{F}(\mathrm{C2T1}) \Rightarrow \#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1})$ can be easily falsified by $\#\mathrm{F}(\mathrm{C1T1}) \wedge \#\mathrm{F}(\mathrm{C2T1}) \wedge \neg\#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1})$, which leads to the [27, 32] incongruence, corresponding to the situation when there are features in both images but there are no tentative matches found which can easily happen for scenes with strong repetitive structures.

Similar reasoning can be used for predicate $\#\mathrm{VMU}(\mathrm{C1T1}, \mathrm{C2T1}) \Rightarrow \#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1})$ which can never be falsified, while the other predicate $\#\mathrm{TM}(\mathrm{C1T1}, \mathrm{C2T1}) \Rightarrow \#\mathrm{VMU}(\mathrm{C1T1}, \mathrm{C2T1})$ can be falsified when there is a lot of tentative matches found but no epipolar geometry could verify them, this situation is incongruent in the sense of [27, 32].

# 5 Incongruence detection in multiple level-of-detail tracking

Another example application of the incongruence framework has been implemented in the area of tracking humans. Although the final goal is a far more

(a) (b) (c)

Figure 17: Sample images representing abnormal situations. (a) Original image. (b) Darkened image. (b) Blurred image.

elaborate scheme with 4 or so levels of tracking, we will here discuss some preliminary results on the basis of just two such levels.

In the course of WP3, tracking modules for analyzing various levels of detail in human motion have been developed. However, as visual data often suffers from noise, small scales, or occlusions, stronger models need to be employed to allow a more detailed analysis. WP3 thus was first concerned with the multi-body blob-level tracking of humans [9, 25], with a focus on extracting independent motions of pedestrians. In this general model, pedestrians are considered blobs that move according to a constant-velocity motion model. Their body pose, or articulation, is dealt with in a more specific tracking model [17, 10] that—due to above stated reasons of noise, as well as spatial and temporal image resolution—assumes periodic movement of the limbs, as is e.g. the case with walking. The two approaches are reviewed in the sections below.

Clearly, these two models fit in with the theory of incongruence detection presented in Section 2. Specifically, we expect the general model to track most kinds of human motion, including erratic ones, while the articulated tracker's specific model might be able to handle the most common movement—walking—but will inevitably fail on any other motion, thus representing an incongruent event according to Table 2 in such a case. An experimental result demonstrating this concept is thus presented in Section 5.3.

## 5.1 Multi-Body Tracking

Our multi-body tracker [9, 25] is based on the same pedestrian detector [6] as used for the audio-visual speaker detector. With the help of Structure-from-Motion, object detections are placed into a common world coordinate system. The actual tracking system then follows a multi-hypotheses approach, where detections of the current and past frames are accumulated in

a space-time volume. This volume is analyzed by growing many trajectory hypotheses using independent bi-directional Extended Kalman filters (EKFs) with a holonomic constant-velocity model.

By starting EKFs from detections at different time steps, an overcomplete set of trajectories is obtained, which is then pruned to a minimal consistent explanation using model selection. This step simultaneously resolves conflicts from overlapping trajectory hypotheses by letting trajectories compete for detections and space-time volume. In a nutshell, the pruning step employs quadratic pseudo-boolean optimization to pick the set of trajectories with maximal joint probability, given the observed evidence over the past frames. This probability

- increases as the trajectories explain more detections and as they better fit the detections' 3D location and 2D appearance through the individual contribution of each detection;
- decreases when trajectories are (partially) based on the same object detections through pairwise corrections to the trajectories' joint likelihoods (these express the constraints that each pedestrian can only follow one trajectory and that two pedestrians cannot be at the same location at the same time);
- decreases with the number of required trajectories through a prior favoring explanations with fewer trajectories – balancing the complexity of the explanation against its goodness-of-fit in order to avoid overfitting ("Occam's razor").

For the mathematical details, we refer to [25]. The most important features of this method are automatic track initialization (usually, after about 5 detections) and the ability to recover from temporary track loss and occlusion.

This method proved to reliably detect independent motions for smooth and even erratic movements, thus constituting our general model of tracking. However, by the ways of the detector [6], it only has the knowledge of the existence of a person, while for a more detailed analysis of the scene, the person's pose would be necessary. As the multi-body tracker is more concerned with data association between world objects and measurements, direct inclusion of body pose would be infeasible. We thus handle it in a separate tracker.

## 5.2 Articulated Tracking

The articulated tracking framework, first introduced in [17], constitutes a stronger model of human locomotion, but is constrained to a single person

and only specific motion patterns.

A generative model is used to learn the regression between a low-dimensional representation of the body pose and the actual pose, the pose and its corresponding image descriptor, as well as the temporal regression between two subsequent poses. Using a particle filtering approach on a low-dimensional manifold, periodic motions such as walking or running can be tracked. Typically, a separate global optimization is performed on the entire sequence to obtain the final result that enforces smooth walking cycles without "switching legs".

The original system relies on silhouettes as image descriptors, as they can be easily obtained in environments with static cameras using background subtraction. For moving cameras, obtaining these silhouettes is more intricate, was however demonstrated in conjunction with the multi-body blob tracker [10]. To decrease dependence on additional processing steps to obtain silhouettes, future versions of this module will investigate other image descriptors such as the HOG descriptor [6] already used throughout several other parts of the project.

While the articulated tracker provides a more informative analysis of human locomotion, it is unable to handle non-periodic or other unknown motions. This is the key to detecting incongruencies, as discussed below.

## 5.3 Incongruence Detection

When applied to an image sequence, both tracking systems will output a probability how well the image fits their world model. For the multi-body tracking, this number is derived from the detection's score, the temporal consistency in appearance (i.e., color histogram similarity between frames), and the smoothness of the trajectory. For the articulated tracker, the probability is obtained from the data likelihood (i.e., similarity of the image descriptor with the predicted model descriptor), as well as the dynamic model.

A sample sequence is shown in Fig. 18. The bounding boxes are the output of the multi-body tracker. The body pose estimated for the person in the front is shown in the lower right, with the articulated tracker's probability shown as a bar next to it (the more red, the less certain the tracker is). At first, both trackers deliver stable performance, as the subject follows a typical walking pattern across the image. However, as the man reaches the street, he startles due to a passing bicycle, thus interrupting the walking cycle. As can be seen, the articulated tracker fails to obtain an suitable explanation and gets stuck at a pose with a low probability, whereas the multi-body tracker successfully continues operation. This thus consistutes an incongruent event according to Table 2, as the general model has a considerably higher proba-

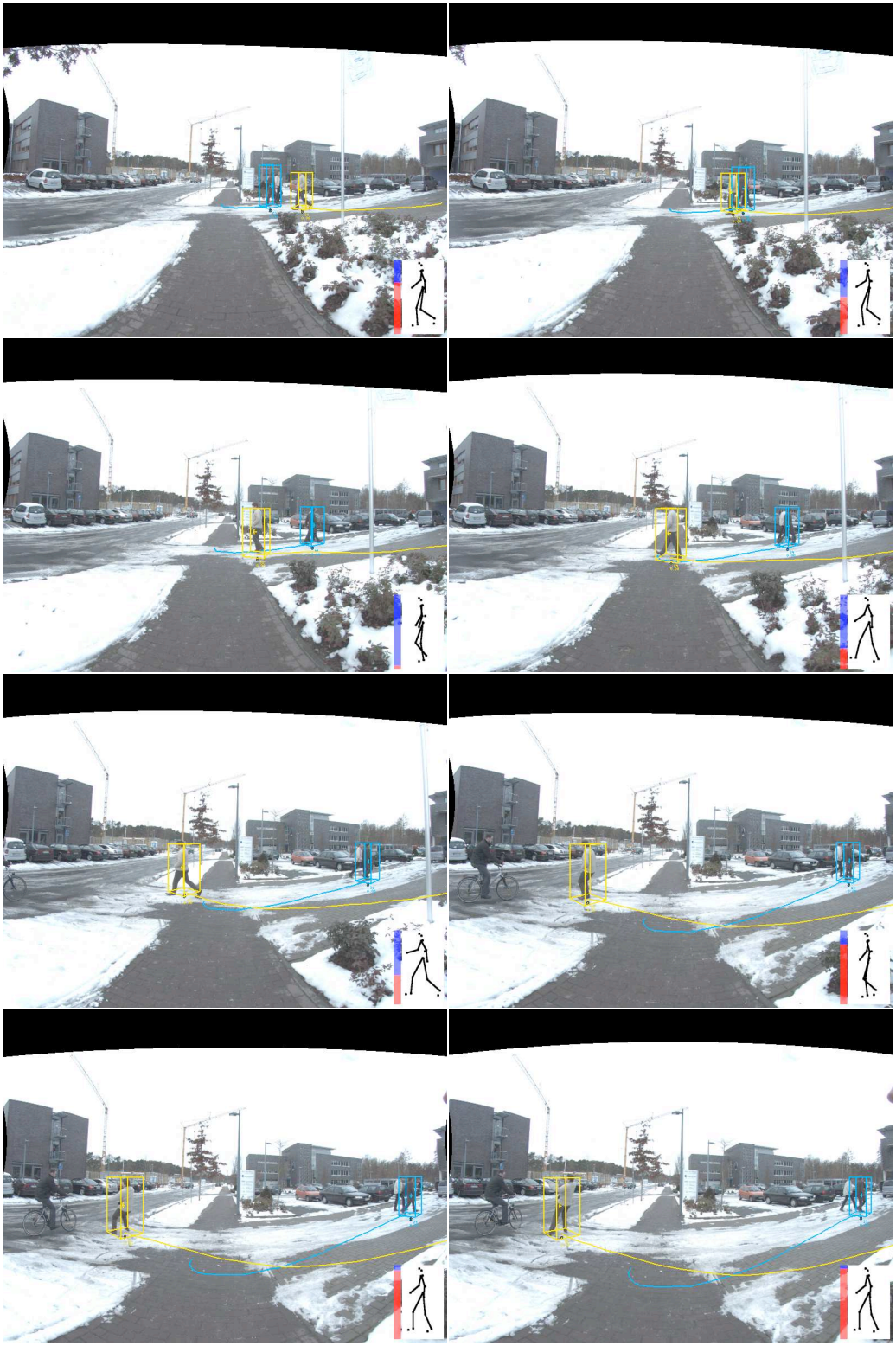


Figure 18: Example tracking result on sequence 0_PedCrossStreet90_2Ped_v2. The yellow bounding box indicates the result of the general classifier (blob tracker), the lower right corner shows the result of the specific classifier (articulated tracker). See text for details.

bility than the specific one. Also note that at the beginning of the sequence, the articulated tracker also reports a comparably low probability due to the partial visibility of other limbs. In some sense, this can also be regarded as incongruency, as this does not occur in the world model of the articulated tracker: it just assumes a single person without any partial occlusions.

Ideally, if such events occur more often, the system will be able to learn such behavior and adapt accordingly, as described in Section 2. As the current tracking system however depends on motion capture data for training, this needs to be investigated further. Furthermore, it will be interesting to explore the opposite case where the general model fails (e.g. due to over-reliance on appearance) whereas the stronger model survives.

# 6 Conclusion

In this research report, we recapitulated and extended the theory of incongruences originally suggested in [32] and showed its application to various parts of the audio-visual processing pipeline envisioned in the DIRAC project. Specifically, we investigated the tasks of "human speaker" detection using either a composite audio-visual classifier or two separate audio and visual classifiers; detecting incongruences on the feature level in camera tracking; and the task of human locomotion analysis by using both a general blob tracker as well as a specific articulated tracker tuned to walking. These preliminary results underline the usefulness of the proposed theory in the application of self-learning and robust systems that should be able to cope with events that are incongruent to their world model.

# References


[1] J. Anemüller, D .Schmidt, and J. Bach. Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features. *Proc. InterSpeech 2008*

[2] H. Bakstein, M. Havlena, P. Pohl, and T. Pajdla. Omnidirectional sensors and their calibration for the Dirac project. Research Report CTU–CMP–2006–13, CMP Prague, December 2006.

[3] M. Barreno, A. A. Cardenas and J. D. Tygar. Optimal ROC Curve for a Combination of Classifiers *NIPS 2007*.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] M. Büchler, S. Allegro, S. Launer, and N. Dillier. Sound classification in hearing aids inspired by auditory scene analysis. EURASIP Journal on Applied Signal Processing, 18, pp. 2991–3002, 2005

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[7] Dirac-CMPdata-19-Awear7. `http://cmp.felk.cvut.cz/projects/dirac/data/Dirac-CMPdata-19.html`.

[8] R. Drullman, J.M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech recognition. J. Acoust. Soc. Am., 95(2), 1994

[9] A. Ess, B. Leibe, K. Schindler, L. Van Gool. A Mobile Vision System for Robust Multi-Person Tracking. *CVPR 2008*.

[10] S. Gammeter, A. Ess, T. Jaeggli, B. Leibe, K. Schindler, and L. van Gool. Articulated multi-body tracking under egomotion. In *to appear in ECCV*, 2008.

[11] S. Greenberg and B. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. Proc. ICASSP, 1997

[12] P. R. Halmos. *Lectures on Boolean Algebras*. Springer, 1974.

[13] M. Havlena, A. Torii, W. Moreau, and T. Pajdla. Omnidirectional audio-visual data acquisition and processing. Research Report CTU–CMP–2008–26, CMP Prague, December 2008.

[14] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd ed., 2004.

[15] D. R. Hofstadter *Gödel, Escher, Bach* Basic Books 1999.

[16] Inria OLT. `http://pascal.inrialpes.fr/soft/olt/`.

[17] T. Jaeggli, E. Koller-Meier, and L. van Gool. Learning generative models for monocular body pose estimation. In *ACCV*, 2007.

[18] T. Joachims Making large-Scale SVM Learning Practical *Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.)*, MIT-Press, 1999.

[19] N. Kanedera, T .Arai, H. Hermansky, and M. Pavel. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication, 28, pp. 43–55, 1999

[20] E. Kidron, Y. Schechner and M. Elad. Pixels that Sound. *CVPR 2005*.

[21] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas. On combining classifiers. *IEEE PAMI 20(3):226-239, 1998*.

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

[23] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 320–327, 1976.

[24] T. S. Kuhn. *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, 1970.

[25] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled detection and tracking from static cameras and moving vehicles. 30(10):1683–1698, 2008.

[26] M. Ostendorf, V. Hohmann, and B. Kollmeier. Classification of acoustical signals based on the analysis of modulation spectra for the application in digital hearing aids. Proc. DAGA, pp. 402–403, 1998

[27] M. Pavel, H. Jimison, D. Weinshall, A. Zweig, F. Ohl, H. Hermansky. Detection and Identification of Rare Incongruent Events in Cognitive and Engineering Systems. *DIRAC White Paper*. DIRAC 2 April 2008.

[28] K. R. Popper. *The Logic of Scientific Discovery*. Routledge, London and New York, 1995.

[29] D. Schmidt and J. Anemüller. Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features. Proc. Interspeech, pp. 2582–2585, 2008.

[30] A. Torii, M. Havlena, M. Jančošek, Z. Kúkelová, and T. Pajdla. Dynamic 3D scene analysis from omni-directional video data. Research Report CTU–CMP–2008–25, CMP Prague, December 2008.

[31] A. Torii, M. Havlena, and T. Pajdla. Omnidirectional image stabilization by computing camera trajectory. In *PSIVT 2009*, pages 71–82, 2009.

[32] D. Weinshall et al. Beyond Novelty Detection: Incongruent Events, when General and Specific Classifiers Disagree. *NIPS 2008*.

[33] E. Zwicker. Subdivision of the audible frequency range into critical bands. J. Acoust. Soc. Am., 33, p. 248, 1961