Incongruence Detection in Audio-Visual Processing

Michal Havlena, Jan Heller, Hendrik Kayser, Jörg-Hendrik Bach, Jörn Anemüller, and Tomáš Pajdla

Abstract The recently introduced theory of incongruence allows for detection of unexpected events in observations via disagreement of classifiers on specific and general levels of a classifier hierarchy which encodes the understanding a machine currently has of the world. We present an application of this theory, a hierarchy of classifiers describing an audio-visual speaker detector, and show successful incongruence detection on sequences acquired by a static as well as by a moving AWEAR 2.0 device using the presented classifier hierarchy.

1 Theory of Incongruence

The recently introduced theory of incongruence [6, 9] allows for detection of unexpected events in observations via disagreement of classifiers on specific and general levels of a classifier hierarchy which encodes the understanding a machine currently has of the world. According to [6, 9], there are two possibilities of how an incongruence can appear. In a class-membership hierarchy, incongruence appears when a general classifier accepts but all the specific, i.e. child, classifiers reject. This means that possibly a novel sub-class of objects has been observed. On the other hand, in a part-whole hierarchy, incongruence appears when all general, i.e. parent, classifiers accept but a specific classifier does not. This is often caused by the fact that the

Michal Havlena · Jan Heller · Tomáš Pajdla
Center for Machine Perception, Department of Cybernetics, FEE, CTU in Prague,
Technická 2, 166 27 Prague 6, Czech Republic
e-mail: {havlem1, hellej1, pajdla}@cmp.felk.cvut.cz

Hendrik Kayser · Jörg-Hendrik Bach · Jörn Anemüller
Medizinische Physik, Fakultät V, Carl von Ossietzky - Universität Oldenburg,
D-26111 Oldenburg, Germany
e-mail: {hendrik.kayser, j.bach, joern.anemueller}@uni-oldenburg.de

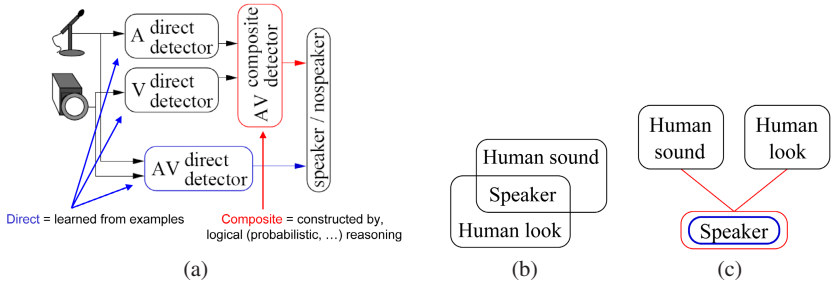


Fig. 1 (a) "Speaker" event can be recognized in two ways, either by a holistic (direct) classifier, which is trained directly from complete audio-visual data, or by a composite classifier, which evaluates the conjunction of direct classifiers for "Human sound" and "Human look" events. (b) "Speaker" is given by the intersection of "Human sound" and "Human look". (c) "Speaker" corresponds to the infimum in the Boolean POSET.

hierarchy is incomplete, some of the classifiers on the general level of the hierarchy may be missing.

Our audio-visual speaker detector is an example of a part-whole hierarchy. Figure 1 shows "Speaker" event that is recognized in two ways, either by the direct classifier, which is trained directly from complete audio-visual data, or by the composite classifier that evaluates the conjunction of direct classifiers for "Human sound" and "Human look" events. "Speaker" is given by the intersection of sets representing "Human sound" and "Human look" which corresponds to the infimum in the Boolean POSET. In the language of [6, 9], 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 Figure 2(a), detects human sound, e.g. speech, and returns a boolean decision on "Human sound" event. The direct visual classifier, see Figure 2(b), detects human body shape in an image and returns a boolean decision on "Human look" event. The direct audio-visual classifier, see Figure 3(a), detects the presence of a speaker and returns a boolean decision on "Speaker" event. The composite audio-visual classifier, see Figure 3(b), constructed as the conjunction of the direct audio and visual classifiers also detects the presence of a speaker and returns a boolean decision on "Speaker" event. Opposed to the direct audio-visual classifier, its decisions are constructed from the decisions of the separate classifiers using logical AND. After presenting a scene with a silent person and speaking loudspeaker, the composite audio-visual classifier accepts but the direct audio-visual classifier does not. That creates a disagreement, incongruence, between classifiers. Table 1 interprets the results of "Speaker" event detection.

According to the theory, collected incongruent observations can be used to refine the definition of a speaker used by the machine, i.e. to correct the understanding of the world, by adding a classifier to the general level of the hierarchy. The composite audio-visual classifier of the refined "Speaker" event detector will not accept scenes

Table 1 Interpretation of agreement/disagreement for "Speaker" event example.

		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

with silent people and speaking loudspeakers because they will be rejected by the added classifier. As the direct audio-visual classifier also rejects such scenes, no incongruence will be rendered. The detailed description of the refinement of the definition of a speaker, which is beyond the scope of this paper, can be found in [5] or in an accompanied paper.

2 Audio-Visual Speaker Detector

Next, we describe the classifiers used in deeper detail.

2.1 Direct Audio Detector

The direct audio classifier for sound source localization is based on the generalized cross correlation (GCC) function [4] of two audio input signals. A sound source is localized by estimating the azimuthal angle of the direction of arrival (DOA) relative to the sensor plane defined by two fronto-parallel microphones, see Figure 2(a). The approach used here enables the simultaneous localization of more than one sound source in each time-frame.

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:

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

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. We use PHAse Transform (PHAT) weighting [4], which normalizes the amplitudes of the input signals to unity in each frequency band:

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

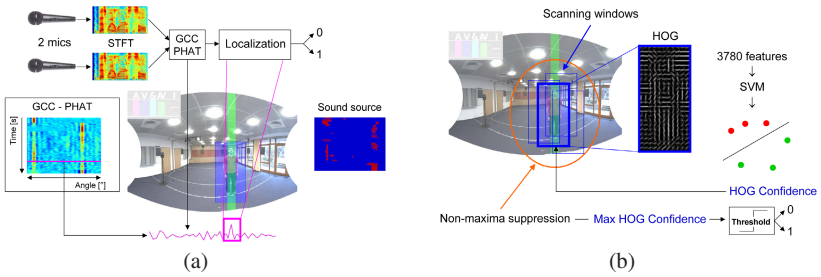


Fig. 2 (a) The direct audio detector employing GCC-PHAT. A sound source is localized by estimating the azimuthal angle of the direction of arrival (DOA) relative to the sensor plane. (b) The direct visual detector based on work [2].

such that only the phase difference between the input signals is preserved.

From the GCC data, activity of a sound source is detected for 61 different DOA angles. The corresponding time delays between the two microphones cover the field of view homogeneously with a resolution of 0.333 ms. The mapping from the time delay τ to the angle of incidence θ is non-linear:

$$\theta = \arcsin\left(\tau \cdot \frac{c}{d}\right), \tag{3}$$

where c denotes the speed of sound and d the distance between the sensors. This results in a non-homogenous angular resolution in the DOA-angle space, with higher resolution near the center and lower resolution towards the edges of the field of view. If at least in one of these 61 directions a sound source is detected, the direct audio classifier accepts.

2.2 Direct Visual Detector

The state-of-the-art paradigm of visual human detection [2] is to scan an image using a sliding detection window technique to detect the presence or absence of a human-like shape. By adopting the assumption that the ground plane is parallel to the image, which is true for our static camera scenario, we can restrict the detection window shapes to those that correspond to reasonably tall (150–190 cm) and wide (50–80 cm) pedestrians standing on the ground plane. Visual features can be computed in each such window as described below.

For the direct visual classifier, we use INRIA OLT detector toolkit [1], based on the histograms of oriented gradients (HOG) algorithm presented by Dalal and Triggs [2]. It uses a dense grid superimposed over a detection window to produce a 3,780 dimensional feature vector. A detection window of 64×128 pixels is divided into cells of 8×8 pixels, each group of 2×2 cells is then integrated into overlapping

blocks. Each cell consists of 9-bin HOG that are concatenated for each block and normalized using L_2 norm. Each detection window consists of 7×15 blocks resulting in 3,780 features. As the cylindrical projection used to represent image data in our experiments is locally similar to the perspective projection, we used a detector trained on perspective images.

An image is scanned by the detection window at several scales and positions in a grid pattern. Detection scores (confidences) higher than 0, i.e. tentative human detections, are further processed using non-maxima suppression with robust mean shift mode detection, see Figure 2(b). If the detection score of the best detection is higher than 0.1, the direct visual classifier accepts.

2.3 Direct Audio-Visual Detector

For the direct audio-visual classifier, we use the concept of angular (azimuthal) bins that allows for handling multiple pedestrians and/or sound sources at the same time. 180° field of view is divided into twenty bins, each 9° wide, with the classification performed per bin. If at least one of the bins is classified as positive, the direct audio-visual classifier accepts. The detailed procedure of the classification of a single bin follows.

First, a 2D feature vector is constructed from audio and visual features as the highest GCC-PHAT value and the highest pedestrian detection score belonging to the bin. The pedestrian detection score is maximized both in the x and y coordinates of the window center where x has to lie in the bin and y goes through the whole height of the image as the bins are azimuthal, i.e. not bounded in the vertical direction. Then, the feature vector is classified by SVM with the RBF kernel [7]. Any non-negative SVM score yields a positive classification.

The SVM classifier was trained using four sequences (2,541 frames in total) of people speaking while walking along a straight line. Every frame of the training sequences contains one manually labeled positive bin and the 2D feature vectors corresponding to these bins were taken as positive examples. The two bins adjacent to the labeled ones were excluded from training and the feature vectors corresponding to the rest of the bins were taken as negative examples. As examples of people not speaking and of speech without people, which were not in the training data, were also needed, we created negative examples for these cases by combining parts of the positive and negative feature vectors, yielding 109,023 negative examples in total. The training data together with the trained decision curve can be seen in Figure 3(a).

2.4 Composite Audio-Visual Detector

The composite audio-visual classifier was constructed to explain audio-visual events by a combination of simpler audio and visual events. At the moment, the classifiers

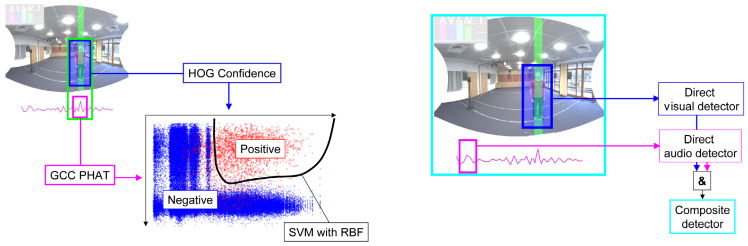


Fig. 3 (a) The direct audio-visual detector constructed from audio and visual features. Training data and the trained decision curve for SVM with RBF of the direct audio-visual classifier: GCC-PHAT values (x-axis) and pedestrian detection scores (y-axis) for different positive (*red circles*) and negative (*blue crosses*) manually labeled examples. (b) The composite audio-visual detector constructed as the conjunction of the decisions of the direct audio and visual detectors.

are combined using a two-state logic, which is too simplistic to cope with the full complexity of real situations but it provides all basic elements of reasoning that later could be modeled probabilistically. Opposed to the direct audio-visual classifier, which separates relevant events from the irrelevant ones, the composite audio-visual classifier represents the understanding a machine currently has of the world.

The composite classifier is not meant to be perfect. It is constructed as the conjunction of the decisions of the direct audio and visual classifiers, each evaluated over the whole image. Hence, it does not capture spatial co-location of the audio-visual events. It detects events concurrent in time but not necessarily co-located in the field of view, Figure 3(b). It will fire on sound coming from a loudspeaker on the left and a silent person standing on the right when that happens at the same moment. This insufficiency is intentional. It models current understanding of the world, which is due to its complexity always only partial, and leads to detection of incongruence w.r.t. the direct audio-visual classifier when human sound and look come from dislocated places.

3 Experimental Results

The direct and composite audio-visual classifiers were used to process several sequences acquired by the AWEAR 2.0 device [3], comprised of 3 computers, 2 cameras with fish-eye lenses heading forwards, and a Firewire audio capturing device with 2 forward and 2 backward-heading microphones. The device can be placed on a table or worn as a backpack thanks to 4 lead-gel batteries which ensure the autonomy of about 3 hours. Audio and visual data are synchronized using a hardware trigger signal.



(a) (b)

Fig. 4 Two example frames from SPEAKER&LOUDSPEAKER sequence showing (a) a congruent and (b) an incongruent situation when the speaker and the loudspeaker sound respectively. Decisions from different direct classifiers are drawn into the frames: direct audio classifier (*magenta*), direct visual classifier (*blue*), direct audio-visual classifier (*green*). The bars in the top-left corner show also the composite audio-visual classifier decisions (*cyan*) and incongruence/wrong model (*red/yellow*).

3.1 Static Device

The 30 s long sequence SPEAKER&LOUDSPEAKER shot at 14 fps acquired by the static device contains a person speaking while walking along a roughly straight line. After a while the person stops talking and the loudspeaker sounds up, which renders an incongruent observation and should be detected.

As the video data come from the left camera of a stereo rig with a 45 cm wide baseline, there is a discrepancy between the camera position and the apparent position of a virtual listener, located at the center of the acquisition platform, to which the GCC-PHAT is computed. To compensate for this, the distance to the sound source and the distance between the virtual listener and the camera need to be known. The listener–camera distance is 22.5 cm and can be measured from the known setup of the rig. The distance to the sound source is assumed to be 1.5 m from the camera. The corrected angle can be then trivially computed from the camera–listener–sound source triangle using a line–circle intersection. Due to this angle correction, the accuracy of incongruence detection is lower for speakers much further away than the assumed 1.5 m. Two example frames from the sequence can be found in Figure 4 together with the decisions from different classifiers.



(a) (b) (c)

Fig. 5 Pedestrian detection in frames stabilized w.r.t. the ground-plane transferred back to the original frame. (a) Non-stabilized frame. (b) Stabilized frame with pedestrian detection. (c) Non-stabilized frame with the transferred pedestrian detection. Notice the non-rectangular shape of the detection window after the transfer.

3.2 Moving Observer

The script of the 73 s long 14 fps sequence MOVINGOBSERVER is similar to the one of SPEAKER&LOUDSPEAKER sequence, the difference being the fact that the AWEAR 2.0 device was moving and slightly rocking when acquiring it. As the microphones are rigidly connected with the cameras, the relative audio-visual configuration remains the same and movement does not cause any problems to the combination of the classifiers. On the other hand, the direct visual classifier is able to detect upright pedestrians only, therefore we perform pedestrian detection in frames stabilized w.r.t. the ground-plane [8] and transfer the results back into the original frames yielding non-rectangular pedestrian detections, see Figure 5.

As the x coordinate of the mouth can be different than the x coordinate of the center of the detected window for non-upright pedestrians, we use the position of the center of the upper third of the window (i.e. the estimated mouth position) in the direct audio-visual classifier instead.

4 Conclusions

We presented an application of the theory of incongruence, a hierarchy of classifiers describing an audio-visual speaker detector, and showed successful incongruence detection on sequences acquired by the AWEAR 2.0 device. Collected incongruent observations can be used to refine the definition of a speaker used by the machine, i.e. to correct the understanding of the world, by adding a classifier to the general level of the hierarchy, see [5] or an accompanied paper for details.

Acknowledgements This work was supported by the EC project FP6-IST-027787 DIRAC and by Czech Government under the research program MSM6840770038. Any opinions expressed in this paper do not necessarily reflect the views of the European Community. The Community is not liable for any use that may be made of the information contained herein.

References

1. Dalal, N.: INRIA Object Detection and Localization Toolkit (2008). Software available at http://pascal.inrialpes.fr/soft/olt
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR 2005, vol. 2, pp. 886–893 (2005)
3. Havlena, M., Ess, A., Moreau, W., Torii, A., Jančošek, M., Pajdla, T., Van Gool, L.: AWEAR 2.0 system: Omni-directional audio-visual data acquisition and processing. In: EGOVIS 2009: First Workshop on Egocentric Vision, pp. 49–56 (2009)
4. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing **24**(4), 320–327 (1976)
5. Pajdla, T., Havlena, M., Heller, J., Kayser, H., Bach, J.H., Anemüller, J.: Incongruence detection for detecting, removing, and repairing incorrect functionality in low-level processing. Research Report CTU–CMP–2009–19, Center for Machine Perception, K13133 FEE Czech Technical University (2009)
6. Pavel, M., Jimison, H., Weinshall, D., Zweig, A., Ohl, F., Hermansky, H.: Detection and identification of rare incongruent events in cognitive and engineering systems. Dirac white paper, OHSU (2008)
7. Schölkopf, B., Smola, A.: Learning with Kernels. The MIT Press, MA (2002)
8. Torii, A., Havlena, M., Pajdla, T.: Omnidirectional image stabilization by computing camera trajectory. In: PSIVT 2009, pp. 71–82 (2009)
9. Weinshall, D., et al.: Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In: NIPS 2008, pp. 1745–1752 (2008)