IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS), London (UK), 2007.

CASSANDRA: audio-video sensor fusion for aggression detection*

W. Zajdel* J.D. Krijnders† T. Andringa† D.M. Gavrila*

* Intelligent Systems Laboratory, Faculty of Science, University of Amsterdam
† Auditory Cognition Group, Artificial Intelligence, Rijksuniversiteit Groningen

Abstract

This paper presents a smart surveillance system named CASSANDRA, aimed at detecting instances of aggressive human behavior in public environments. A distinguishing aspect of CASSANDRA is the exploitation of the complimentary nature of audio and video sensing to disambiguate scene activity in real-life, noisy and dynamic environments. At the lower level, independent analysis of the audio and video streams yields intermediate descriptors of a scene like: "scream", "passing train" or "articulation energy". At the higher level, a Dynamic Bayesian Network is used as a fusion mechanism that produces an aggregate aggression indication for the current scene. Our prototype system is validated on a set of scenarios performed by professional actors at an actual train station to ensure a realistic audio and video noise setting.

1 Introduction

Surveillance technology is increasingly fielded to help safeguard public spaces such as train stations, shopping malls, street corners, in view of mounting concerns about public safety. Traditional surveillance systems require human operators who monitor a wall of CCTV screens for specific events that occur rarely. Advanced systems have the potential to automatically filter-out spurious information and present the operator only the security-relevant data. Existing systems have still limited capabilities; they typically perform video-based intrusion detection, and possibly some trajectory analysis, in fairly static environments.

In the context of human activity recognition in dynamic environments, we focus on the relatively unexplored problem of aggression detection. Earlier work involved solely the video domain and considered fairly controlled in-door environments with static background and few (two) persons [2]. Because events associated with the build-up or enactment of aggression are difficult to detect by a single sensor modality (e.g. shouting versus hitting-someone), in this work we combine audio- and video-sensing. At the low level raw sensor data are processed to compute 'intermediate-level" events or features that summarize activities in the scene. Examples of such descriptors implemented in the current system are "scream" (audio) or "train passing" and "articulation energy" (video). At the top level, a Dynamic Bayesian Network combines the visual and auditory events and incorporates any context-specific knowledge in order to produce an aggregate aggression indication. This is unlike previous work where audio-video fusion dealt with speaker localization for advanced user-interfaces, i.e. using video information to direct a phased-array microphone configuration [8].

2 System Description

2.1 Audio Unit

Audio processing is performed in the time-frequency domain with an approach common in auditory scene analysis [11]. The transformation of the time-signal to the time-frequency domain is performed by a model[1] of the human ear [3]. This model is a transmission-line model with its channels tuned according to the oscillatory properties of the basilar membrane. Leaky-integration of the squared membrane displacement results in an energy-spectrum, called a cochleogram (see Fig. 1).

A signal component is defined as a coherent area of the cochleogram that is very likely to stem from a single source. To obtain signal components, the cochleogram is first filtered with a matched filter which encodes the response of the cochlea to a perfect sinusoid of the frequency applicable to that segment. Then the cochleogram is thresholded using a cut-off value based on two times the standard deviation of the energy values. Signal components are obtained as the tracks formed by McAulay-Quatari tracking [7], applied on this pre-processed version of the cochleogram. This entails stringing the energy maxima of connected components together, over the successive frames.

*This research was supported by the Dutch Science Foundation NWO under grant 634.000.432 within the ToKeN2000 program.

[1]We thank Sound Intelligence (http://www.soundintel.com) for contributing this model and cooperation on the aggression detection methods.

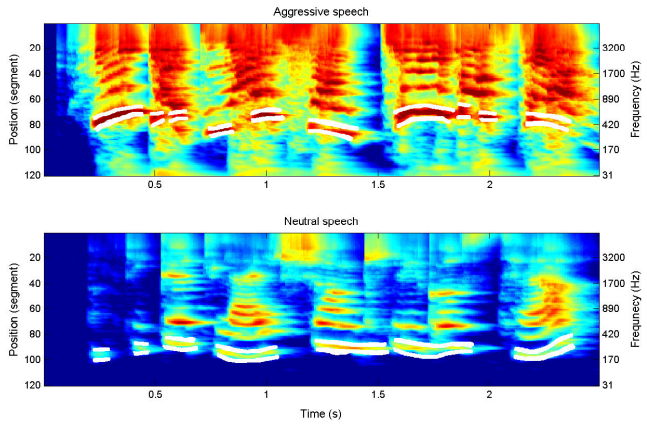

Figure 1: Typical cochleograms of aggressive and normal speech (top and bottom figure, respectively). Energy content is color-coded (increasing from blue to red). Note the higher pitch (marked by white lines) and the more pronounced higher harmonics for the aggressive speech case.

Codeveloping sinusoids with a frequency development equal to an integer multiple of a fundamental frequency (harmonics) are subsequently combined into harmonic complexes. Note that these harmonics can be combined safely because the probability is small that uncorrelated sound sources show this measure of correlation by chance.

Little or no literature exists on the influence of aggression on the properties of speech. However the Component Process Model from Scherer [9] and similarities with the Lombard reflex [6] suggest a couple of important cues for aggression. The component process theory assumes that anger and panic, emotions strongly related to aggression, are seen as a ergotropic arousal. This form of arousal is accompanied by an increase in heart frequency, blood pressure, transpiration and associated hormonal activity. The predictions given by the model show many similarities with the Lombard reflex.the vocal chords increases the pitch and enhances the higher harmonics, which leads to an increase in spectral tilt. These properties, pitch (fundamental frequency (f_0)) and spectral tilt (a measure of the slope of the average energy distribution, calculated as the the energy of the harmonics above 500 Hz divided by the energy of the harmonics below 500 Hz) are calculated from the harmonic complexes. The audio detector uses these two properties as input for a decision tree. An example of normal and aggressive speech can be seen in Fig. 1.

2.2 Video Unit

Analysis of the video stream aims primarily at computing visual cues characteristic for physical aggression among humans. Physical aggression is usually characterized by fast articulation of body parts (i.e. arms, legs). Therefore, a principled approach for detecting aggression involves detailed body-pose estimation, possibly in 3D, followed by ballistic analysis of movements of body parts. Unfortunately, at present pose estimation remains a significant computational challenge. Various approaches [4] operate at limited rates and handle mostly a single person in a constrained setting (limited occlusions, pre-fitted body model).

Simplified approaches rely on a coarser representation human body. An example is a system [2] that tracks a head of a person by analyzing body contour and correlates aggression with head's "jerk" (derivative of acceleration). In practice, high-order derivatives related to body contours are difficult to estimate robustly in cases where the background is not static and there is a possibility of occlusion.

2.2.1 Visual aggression features

Here we consider alternative cues based on an intuitive observation that aggressive behavior leads to highly energetic body articulation. We estimate (pseudo-) kinetic energy of body parts using a "bag-of-points" body model. The approach relies on simple image processing operations and yields features highly correlated with aggression.

For detecting people our video subsystem employs adaptive background/foreground subtraction technique [12]. The assumption of static background scene holds fairly well in the center view area, where most of the people enter the scene. After detection, people are represented as ellipses and tracked with an extended version of the mean-shift tracking algorithm [13]. The extended tracker adapts position *and* shape of the ellipse (tilt, axes) and thus facilitates a close approximation of body area even for tilted/bended poses. Additionally, the mean-shift tracker handles well partial occlusions and achieves near real-time performance.

We consider human body as a collection of loosely connected points with identical mass. While such a model is clearly a simplification, it reflects well the non-rigid nature of a body and facilitates fast computations. Assuming Q points attached to various body-parts, the average kinetic energy of an articulating body is given by the average kinetic energy of points,

$$E = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{2}m_i|v_i - v_\mathrm{e}|^2 \qquad (1)$$

where v_i, m_i denote, respectively, velocity vector and mass of the ith point, and v_e denotes the velocity vector of the ellipse representing the person. By discounting overall body motion we capture only the articulation energy.

Due to the assumption of uniform mass distribution between points, the total body mass becomes a scale factor. By omitting scale factors, we obtain a *pseudo*-kinetic energy estimate in the form $\bar{E} = \frac{1}{Q}\sum_{i=1}^{Q}|v_i - v_\mathrm{e}|^2$. Such

Figure 2: (Left) Optical-flow features for detecting trains in motion. (Right) Representing people: ellipses (for tracking) and points (for articulation features).

features are assumed to measure articulation and will be our primary visual cues for aggression detection.

Computation of energy features requires selecting image points that represent a person. Ideally, the points would cover the limbs and the head since these parts are mostly articulated. Further, to estimate velocities, the selected points must be easy to track. Accordingly, we select $Q = 100$ points within an extended bounding box of a person by finding pixels with the most local contrast [10]. Such points are easy to track and usually align well with edges in an image (which in turn often coincide with limbs as in Fig. 2, right). For point tracking we use the KLT algorithm [10] (freely available implementation from the OpenCV library).

2.2.2 Train detection

An additional objective of the video unit is detecting trains *in motion*. Trains moving in and out of a station produce auditory noise that often leads to spurious audio-aggression detections. Therefore recognizing trains in video opens a possibility for later suppressing of such detections.

A train usually appears as a large, rigid body and moves along a constrained trajectory. For a given view and rail section we define a mask that indicates the image regions where a train typically appears. In this region we track frame-to-frame motion of $N = 100$ image features with KLT [10] tracker (Fig. 2, left). The features' motion vectors are classified as train/non-train by a pre-trained nearest-neighbor classifier. A train in motion is detected when more then 50% of the features are classified positively. Due to the constrained movement of trains, our detector turns out quite robust to occasional occlusions of the train area by people.

2.3 Fusion Unit

The fusion unit produces an aggregate aggression indication given the features/events produced independently by the audio and video subsystems. A fundamental difficulty with fusion arises from inevitable ambiguities in human behavior which make it difficult to separate normal from aggressive activities (even for a human observer). Additional problems follow from various noise artifacts in the sensory data.

Given the noisy and ambiguous domain we resort to a probabilistic formulation. The fusion unit employs a probabilistic time-series model (a Dynamic Bayesian Network, DBN [5]), where aggression level can be estimated in a principled way by solving appropriate inference problem.

2.3.1 Basic model

We denote the discrete-time index as $k = 1, 2, \ldots$, and set the gap between discrete-time steps (clock ticks) to 50 ms. At the kth step, $y^{\mathrm{a}}_k \in \{0, 1\}$ denotes the output of audio aggression detector, and $y^{\mathrm{v}}_{j,k}$ denotes the pseudo-kinetic energy of the jth, $j = 1, \ldots, J$, person. Our system can comprise several train detectors monitoring non-overlapping rail sections. The binary output of the mth, $m = 1, \ldots, 4 = M$, train detector will be denoted as $y^{\mathrm{T}}_{m,k} \in \{0, 1\}$. (We tested a configuration with $M = 4$.)

Our aim is to detect "ambient" scene aggression, without deciding precisely which persons are aggressive. Therefore we reason on the basis of a cumulative articulation measurement $y^{\mathrm{v}}_k = \sum_j y^{\mathrm{v}}_{j,k}$ over all persons. Additionally, the cumulative is quite robust to (near-)occlusions when articulation of one person could be wrongly attributed to another.

In order to reason about aggression level, we use a 5-step discrete scale $\langle 0, 1 \rangle$: 0.0 (no activity), 0.2 (normal activity), 0.4 (attention required), 0.6 (minor disturbance), 0.8 (major disturbance), and 1.0 (critical aggression).

Importantly, the aggression level obeys specific correlations over time and should be represented as a process (rather than an instantaneous quantity). We will denote aggression level at step k as a_k and define a stochastic process $\{a_k\}$ with dynamics given by a 1st order model:

$$p(a_{k+1} = i | a_k = j) = \mathrm{CPT}_a(i, j), \qquad (2)$$

where $\mathrm{CPT}_a(i, j)$, denotes a conditional probability table. In a sense, the first order model is a simplification as it captures only short-term dependencies.

The measured visual (y^{v}_k) and auditory (y^{a}_k) features are treated as samples from an observation distribution (model) that depends on the aggression level a_k. Since (later on) we will incorporate information about passing trains, we introduce a latent train-noise indicator variable $n_k \in \{0, 1\}$ and assume that the observation model

$$p(y^{\mathrm{v}}_k, y^{\mathrm{a}}_k | a_k, n_k)$$

depends also on the train-noise indicator. The model takes the form of a conditional probability table CPT_o, where the cumulative articulation feature is discretized.

2.3.2 Train models

The fusion DBN comprises several subnetworks — train models which couple train detections $y^{\mathrm{T}}_{m,k}$ with the latent

train-noise indicator n_k. Additionally, each train model encodes prior information about duration of a train pass.

For the mth rail section, we introduce a latent indicator $i_{m,k} \in \{0, 1\}$ of a train passing at step k. We assume that the train detections $y^{\mathrm{T}}_{m,k}$, the train-pass indicators $i_{m,k}$, and the train noise n_k obey a probabilistic relation

$$p(y^{\mathrm{T}}_{m,k}|i_{m,k}) = \mathrm{CPT}_t(y^{\mathrm{T}}_{m,k}, i_{m,k}) \quad (3)$$

$$p(n_k|i_{1:M,k}) = \mathrm{CPT}_n(n_k, i_{1:M,k}). \quad (4)$$

For each rail, the model (3) encodes inaccuracies of detector (mis-detections, false alarms). The model (4) represents the fact that passing trains usually induce noise, but also that sometimes noise is present without a passing train.

Since a typical pass takes $5 - 10$ seconds ($100 - 200$ steps) the pass indicator variable exhibits strong temporal correlations. We represent such correlations with a time-series model based on a gamma distribution. A gamma pdf $\gamma(\tau_m, \alpha_m, \beta_m)$ is a convenient choice for modeling duration τ_m of an event (α_m, β_m are parameters). To apply this model in a time-series formulation, we replace the total duration τ_m with a partial duration $\tau_{m,k}$ that indicates how long a train is already passing a scene at step k.

By considering a joint process $\{i_{m,k}, \tau_{m,k}\}$ temporal correlations can be enforced by the following model

$$\begin{aligned} &p(i_{m,k+1} = 1|\tau_{m,k}, i_{m,k} = 0) = \eta_m \\ &p(i_{m,k+1} = 1|\tau_{m,k}, i_{m,k} = 1) = p(\tau_m > \tau_{m,k}) = \\ &= \int_{\tau_{m,k}}^{+\infty} \gamma(\tau_m, \alpha_m, \beta_m)\mathrm{d}\tau_m = 1 - F(\tau_{m,k}, \alpha_m, \beta_m), \end{aligned}$$

where $F()$ is a gamma cumulative density function. Parameter η_m denotes a probability of starting a new train pass. At the kth step, the probability of continuing a pass is function of the current duration of the pass. A configuration ($i_{m,k+1} = 1$, $\tau_{m,k}$, $i_{m,k} = 1$) implies that a pass does not finish yet and the total pass duration will be larger than $\tau_{m,k}$, hence the integration. Further, the partial duration variable obeys a deterministic regime

$$\tau_{m,k+1} = \begin{cases} 0 & \text{iff } i_{m,k+1} = 0 \\ \tau_{m,k+1} = \tau_{m,k} + \epsilon & \text{otherwise} \end{cases},$$

where $\epsilon = 50\,\mathrm{ms}$ is the period between successive steps.

2.3.3 Inference and Learning

In a probabilistic framework, reasoning about aggression corresponds to solving probabilistic inference problems. In an online mode, the key quantity of interest is the posterior distribution on aggression level given data collected at up to the current step, $p(a_k|y^{\mathrm{v}}_{1:k}, y^{\mathrm{a}}_{1:k}, y^{\mathrm{t}}_{1:m,1:k})$. From this distribution we calculate the expected aggression value, which will be the basic output of the fusion system.

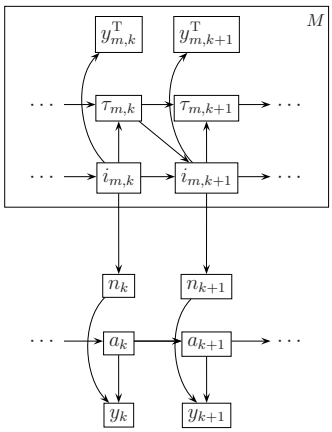

Figure 3: Dynamic Bayesian Network representing the probabilistic fusion model. The rectangular plate indicates $M = 4$ replications of the train sub-network.

Given the graphical structure of the model (Fig. 3), the required distribution can be efficiently computed using a recursive, forward filtering procedure [5]. We implemented an approximate variant of the filtering procedure, known as the Boyen-Koller algorithm [1]. At a given step k, the algorithm maintains only marginal distributions $p(h_k|y^{\mathrm{v}}_{1:k}, y^{\mathrm{a}}_{1:k}, y^{\mathrm{t}}_{1:m,1:k})$, where h_k is any of the latent variables. When new detector data arrive, the current-step marginals are updated to represent the next-step marginals.

An important modeling aspect are temporal developments of processes in the scene. Unlike the binary train-pass events, aggression level usually undergoes more subtle evolutions as the tension and anger among people build up. Since the assumed (1st-order) model might not capture well long-term effects and a stronger model would be rather complicated we enforce temporal correlations with a simple low-pass filter. The articulation measurements (before inference) and the expected aggression level (after inference) are low-pass filtered using a $10\,\mathrm{s}$ running-average filter.

The parameters of our model: probability tables: CPT_a, CPT_o, CPT_n, CPT_t and the parameters α_m, β_m of the gamma pdf's) are estimated by maximum-likelihood learning. The learning process relies on detector measurements from training audio-video clips and ground-truth annotations. The annotations are particularly important for learning the observation model CPT_o. An increased probability of a false auditory aggression in the presence of train noise, will suppress the contribution of audio data to the aggression level when the video subsystem reports a passing train.

3 Experiments

We evaluate the aggression detection system using a set of 13 audio-video clips (scenarios) recorded at a train station.

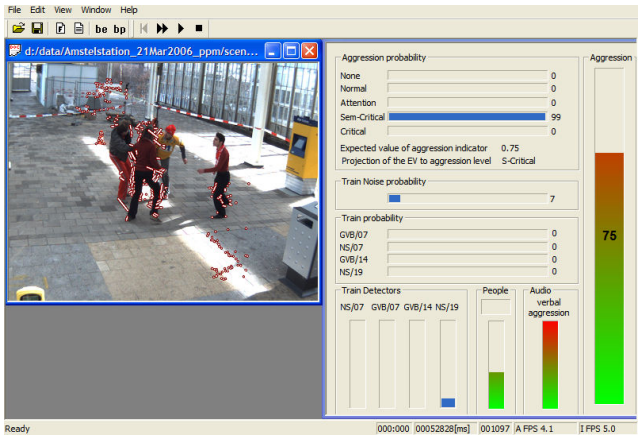

Figure 4: A screen-shot of the CASSANDRA prototype system. In the lower part of the right window, various intermediate quantities are shown: the probabilities for trains on various tracks (currently zero), the output of the video-based articulation energy ("People", currently mid-way), the output of the audio detector (currently at maximum). The slider covering the right side shows the overall estimated aggression level (currently "major disturbance").

The clips (each $100s - 150s$) feature 2-4 professional actors who engage in a variety of activities ranging from normal (walking) through slightly excited (shouting, running, hugging), moderate aggressive (pushing, hitting a vending machine) to critically aggressive (football-supporters clashing). The recording took place at a platform of an actual train station (between two rail tracks, partially outdoor) and therefore incorporates realistic artifacts, like noise and vibrations from trains, variable illumination, wind, etc.

Scenarios have been manually annotated with a ground-truth aggression level by two independent observers using the scale mentioned in Sect. 2.3.1. Aggression toward objects was rated approx. 25% lower than aggression toward humans, i.e. the former did not exceed a level of 0.8.

Fig. 5 details the results of the CASSANDRA system on a scenario involving gradual aggression build-up. Here, two pairs of competing supporters first start arguing, then get in a fight. The bottom panel shows some illustrative frames, with articulation features highlighted. We see from Fig. 5 that the raw (before low-pass filtering) articulation measurements are rather noisy, however the low-pass filtering reveals strong correlation with ground-truth aggression level. The effect of low-pass filtering of the estimated aggression level is shown bottom plot of Fig. 5. A screen-shot of the CASSANDRA system in action is shown in Fig. 4.

We considered two quantitative criteria to evaluate our system. The first is the deviation of the CASSANDRA estimated aggression level from the ground-truth annotation. Here we obtained a deviation of mean 0.17 with standard

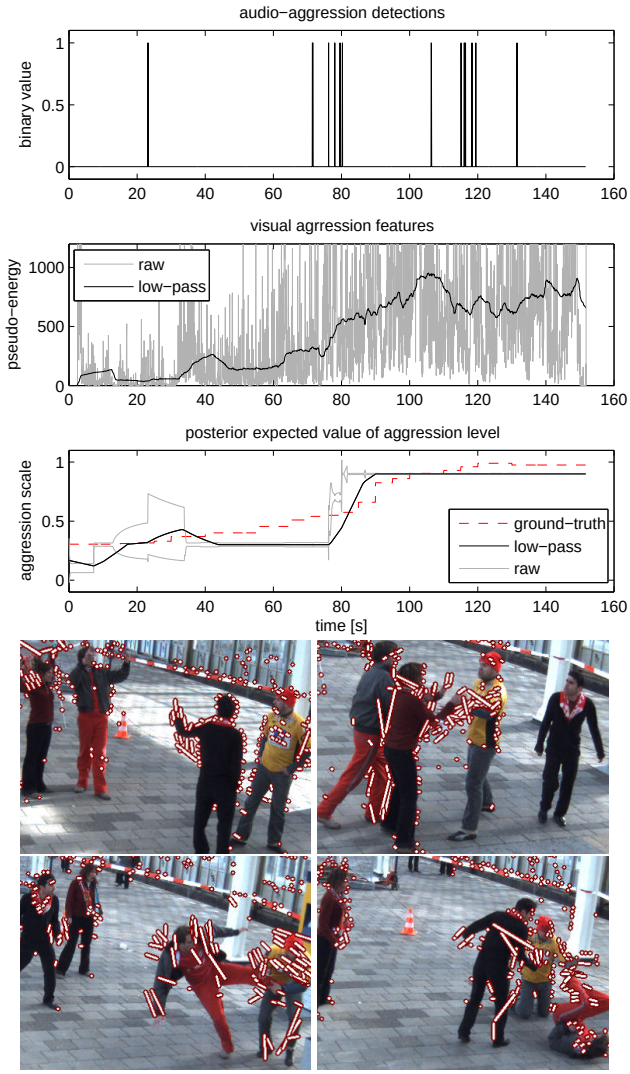

Figure 5: Aggression build-up scenario. (Top graph) Audio-aggression detections. (Middle graph) Visual articulation measurements. (Bottom graph) Estimated and ground-truth aggression level. The gray lines show uncertainty intervals ($2\times$ std. deviation) around raw (before filtering) expected level. (Images) Several frames (timestamps: 45 s, 77 s, 91 s, 95 s). Notice correspondence with the articulation measurements.

Figure 6: Selected frames from a scenario involving aggression toward a machine.

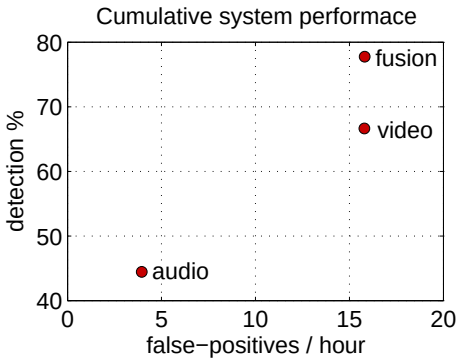

Figure 7: Cumulative detection results by sensor modality.

id	scenario content	ground-truth positive	detected events true-pos.	false-pos.
1	normal: walking, greeting	0	0	0
2	normal: walking, greeting	0	0	1
3	excited: lively argument	0	0	0
4	excited: lively argument	1	1	0
5	aggression toward a vend. machine	1	0	1
6	aggression toward a vend. machine	1	0	0
7	happy football supporters	1	1	0
8	happy football supporters	0	0	1
9	supporters harassing a passenger	1	1	0
10	supporters harassing a passenger	1	1	0
11	two people fight, third intervenes	1	1	0
12	four people fighting	1	1	0
13	four people fighting	1	1	1

Table 1: Aggression detection results by scenario. The table indicates number of events (positive=aggressive). Figure 6 shows example frames from the 5th scenario.

deviation 0.1. The second performance criterion considers aggression detection as a two-class classification problem of distinguishing between "normal" and "aggressive" events (by thresholding aggression level at 0.5). Matching ground-truth with estimated events allows us to compute detection rate ($\%$) and false-alarm rate (per hour). When matching events we allowed a time deviation of 10 s. The cumulative results of leave-one-out tests on 13 scenarios (12 for training, 1 for testing) are given in Fig. 7. Comparing the test results for three modalities (audio, video, fusion of audio+video), we notice that the auditory and visual features indeed are complimentary; with fusion the overall detection rate increased without introducing additional false alarms. It is important to note that our data set is heavily biased toward occurrences of aggression, i.e. which put the system to a difficult test. We expect CASSANDRA to produce much less false alarms in a typical surveillance setting, where most of the time nothing happens.

Table 1 gives an overview of the detection results on the scenarios. We notice that the system performed well on the clearly normal cases (scenarios 1-3) or aggressive cases (scenarios 9-13), while borderline scenarios were more difficult to classify. The borderline behavior (e.g. scenarios 7-8) turns out also difficult to classify for human observers given the inconsistent ground-truth annotation in Tab. 1.

The CASSANDRA system runs on two PCs (one with the video and fusion units, the other with the audio unit). The overall processing rate is approx. 5 Hz with $756 \times 560 \times 20$ Hz input video stream and 44 kHz input audio stream.

4 Conclusions and Future Work

We demonstrated a prototype system that uses a Dynamical Bayesian Network to fuse auditory and visual information for detecting aggressive behavior. On the auditory side, the system relies on scream-like cues, and on the video side, the system uses motion features related to articulation. We obtained a promising aggression detection performance in a complex, real-world train station setting, operating in near-realtime.

The present system is able to distinguish well between clear cases of aggression and normal behavior. In the future, we plan to focus increasingly on the "borderline" cases. For this, we expect to use more elaborate auditory cues (laughter vs scream), more detailed visual cues (indications of body-contact, partial body-pose estimation), and stronger use of context information.

References

[1] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *UAI*, 1998.

[2] A. Datta et al. Person-on-person violence detection in video data. *ICPR*, 01:10433, 2002.

[3] H. Duifhuis et al. *Cochlear Mechanisms: Structure, Function and Models*, pages 395–404. 1985.

[4] D.M. Gavrila. The Visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[5] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *IJPRAI*, 15(1):9–42, 2001.

[6] J.-C. Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1 1993.

[7] R. McAulay and T. Quatieri. Speech analysis / synthesis based on a sinusoidal representation. *Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[8] A. Pentland. Smart rooms. *Scientific American*, 274(4):54–62, 1996.

[9] K. Scherer. Vocal affect expression: A review and a model for future research. *Psychol. Bulletin*, 99(2):143–165, 1986.

[10] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[11] D. Wang and G. Brown. *Computational Auditory Scene Analysis*. John Wiley and Sons, 2006.

[12] Z. Zivkovic. Improved adaptive gausian mixture model for background subtraction. In *ICPR*, 2004.

[13] Z. Zivkovic and B. Krose. An EM-like algorithm for color-histogram-based object tracking. In *CVPR*, 2004.