Computer Vision and Image Understanding 144 (2016) 106–120

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Multi-modal human aggression detection

J.F.P. Kooij [a], M.C. Liem [a], J.D. Krijnders [b,1], T.C. Andringa [b], D.M. Gavrila [a,*]

[a] Intelligent Systems Laboratory, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands
[b] Auditory Cognition Group, Artificial Intelligence, Rijksuniversiteit Groningen, Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 19 December 2014
Accepted 18 June 2015

Keywords:
Automated video surveillance
Multi-modal sensor fusion
Aggression detection
Dynamic Bayesian Network

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 complementary audio and video cues to disambiguate scene activity in real-life environments. From the video side, the system uses overlapping cameras to track persons in 3D and to extract features regarding the limb motion relative to the torso. From the audio side, it classifies instances of speech, screaming, singing, and kicking-object. The audio and video cues are fused with contextual cues (interaction, auxiliary objects); a Dynamic Bayesian Network (DBN) produces an estimate of the ambient aggression level.

Our prototype system is validated on a realistic set of scenarios performed by professional actors at an actual train station to ensure a realistic audio and video noise setting.

1. Introduction

Surveillance cameras are frequently installed to help safeguard public spaces such as train stations, shopping malls, street corners, in view of mounting concerns about public safety. Traditional CCTV systems require human operators to monitor a wall of video screens for specific events that occur rarely. However, due to the large number of video streams and limited human concentration abilities, the chance of an incident actually being noticed may be much lower than one might expect [1]. Smart surveillance systems have the potential to automatically filter-out spurious information and present the operator only the security-relevant data. Most current systems are video-only and limited in their abilities to deal with complex environments containing multiple persons and dynamic backgrounds.

The proposed CASSANDRA[2] system aims to detect human aggression in a complex real-world environment. It combines video and audio cues, together with contextual cues, by means of a Dynamic Bayesian Network to estimate the ambient aggression level in a scene. Fig. 1 shows a screenshot of the system in action. The estimated aggression level is visualized in the large vertical bar at the left; its high value is due to a group of people fighting.

The main visual indicator for physical aggression is fast articulation of body parts (arm swinging, kicking). Ideally, one would perform detailed pose recovery for every person per video frame to accurately estimate body part motion trajectories. But recovering body pose under varying lighting conditions, varying appearances and multiple occlusions is currently still an unsolved problem without a robust and computationally efficient solution. Therefore we aggregate optical flow over a foreground region to capture a person's articulation energy. The multi-view setup can detect 2D motion even when it cannot be clearly seen in some views due to the motion direction or occlusion. The observed motion features are fused per individual across the different camera views with person specific foreground masks. This is achieved by reconstructing the 3D scene with voxel carving and tracking persons in the resulting voxel space.

Even when no physical assault is perceived, the audio signal can contain cues in anticipation of aggression and intimidation, such as shouting. As expected, detecting audio events in real-world environments is challenging due to multiple audio sources, some even located outside an observed scene, and reverberation. CASSANDRA therefore detects and classifies audio events from a preselected set of informative sounds that can still be distinguished from background noise. We show that the combination of auditory and visual aggression cues improves the discriminative power of the system to recognize aggressive situations. Note that while some sound events are characteristic for the enactment of aggression, such as screams or impact sounds when damaging property, other sounds are indicative of non-aggressive situations, such as normal talking. There can also

* Corresponding author.
E-mail address: d.m.gavrila@uva.nl (D.M. Gavrila).

[1] Author is now with the Cognitive Systems Group at INCAS3, Assen, The Netherlands.

[2] In Greek mythology, the daughter of Priam, the last king of Troy, and his wife Hecuba. Cassandra was loved by the god Apollo who promised her the power of prophecy if she would comply with his desires. Cassandra accepted the proposal, received the gift, and then refused the god her favors. Apollo revenged himself by ordaining that her prophecies should never be believed (source: Encyclopedia Britannica).

http://dx.doi.org/10.1016/j.cviu.2015.06.009
1077-3142/

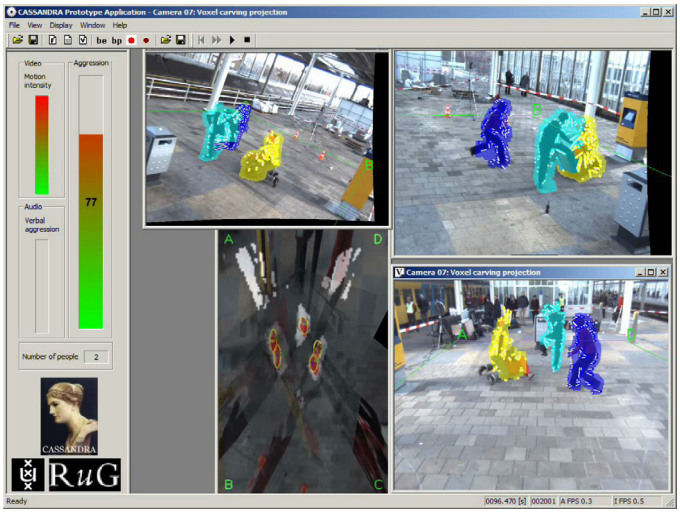

Fig. 1. A screenshot of the CASSANDRA prototype application. In the application three windows show camera images in which detected persons are annotated by a (random) color. A fourth window in the bottom left displays the top-down projection (constructed from image homography) with voxel carving results overlaid in red. On the left side the large vertical bar shows the expected aggression level as computed by the system at that time step. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

be intermediate situations where the interpretation depends on the recording setting. For instance, chanting supporter groups could indicate a tense situation at a generally quiet station, or relatively normal situation (though with some supervision required) near a sports stadium. Since formulating the relation between the various auditory and visual cues to aggression is challenging, we instead estimate model parameters from annotated training data. Such data could either be collected at one particular location for a tailored model, or obtained at various locations for a more general model.

The prototype system is validated on a set of scenarios performed by professional actors at an actual train station to ensure a setting with realistic audio and video noise. The scenarios include multiple persons and person interactions, displaying normal behavior, physical aggression, vandalism, and difficult borderline cases such as loud celebrating football supporters. The train station hallway is a large space with big windows, resulting in naturally changing lighting conditions, shadows and sound reverberation due to the acoustics of the building. It is filled with every day activity such as trains passing by, passengers boarding and exiting carriages, people standing and walking in the background; this makes accurate foreground segmentation quite challenging.

2. Previous work

According to the prevalent definition, “aggression is any form of behavior directed toward the goal of harming or injuring another living being who is motivated to avoid such treatment” [2]. As human aggression is an active field of study in psychology and other social sciences, several attempts have been made to quantify aggression. Most rating scales consist of self-report questionnaires, which ask people about their own experiences and feelings of aggression (e.g. “I sometimes feel very angry”). One of the few to involve observable behavior is the Overt Aggression Scale (OAS) [3]. OAS divides violent behavior in four categories: 1) verbal aggression 2) physical aggression against objects 3) physical aggression against self and 4) physical aggression against other people. Aggressive behavior is rated within each category, guided by some representative examples. Still, rating remains subjective in parts and difficult to assess from direct observations (e.g. distinguishing between minor versus serious injuries).

Given the advanced perceptual and cognitive abilities that are necessary to detect human aggression, and the fluid rating scales, automatic sensor-based aggression detection still stands in its infancy. There is, however, extensive literature on human activity recognition, mainly from a computer vision perspective (see surveys [4–6]). We review this literature by focusing on visual features, audio features, and models for high-level fusion of temporal and contextual data.

2.1. Visual feature extraction

Different image features have been proposed for human activity recognition schemes. Common features for classifying single person activity include Spatio Temporal Interest Points (STIPs) [7], shape-context [8], optical flow [9–12], spatial position and velocity [8,9], Motion Histogram Images [10], and (approximate) body-part positions [13–16]. Visual features can also be learned from large amounts of data directly, e.g. with Convolutional Neural Networks [17], which have been recently applied to video classification too [18]. Motion in particular was found to be a good identifier for overt violence in different applications. In [13] sudden large changes in tracked head positions were used as an indicator of person-on-person violence. And, Hassner et al. [11] showed that analysis of the magnitude changes in optical flow over time can also provide good features to detect overt violence in videos of large crowds. However, measured motion may not only originate from the object of interest, but also from other objects and camera movements, in which case separating foreground motion features from the background improves classification considerably [12].

Various methods have been proposed to combine behavioral observables into activities with a larger temporal extent, such as Petri Nets, (stochastic) context-free grammars and logic-based methods relying on explicit domain knowledge (cf. survey [6]). Typically, long term activity semantics are represented as a latent state that is conditionally dependent on the low level features. Activities can even themselves be combined hierarchically into high-level behaviors patterns [19]. Certain activities are defined in terms of interaction between multiple people, such as walking in a group, ignoring each other, gathering, or fighting. In these cases, single person activity features alone are inadequate [20]. Instead features based on trajectories, such as relative position and relative velocity, have been used to classify observed group activity [20–22].

Recognizing activities of individuals, and/or their relations to others within a group, relies extracting behavioral features per individual, which requires tracking multiple people simultaneously. For fixed viewpoint video surveillance, the classical approach is to track in the image plane, e.g. a standard mean-shift tracker or extract silhouette blobs with background subtraction within a single image (e.g. [9,21,23]). Alternatively, one can track the position of people on the ground plane, since the camera can be intrinsically and extrinsically calibrated [24]. Furthermore, in scenarios with cluttered environments containing (partially) occluded people, complementary observations from the different viewpoints can improve robustness over single-view tracking. The tracked ground plane position of an individual is then a convenient view—invariant representation for subsequent behavior modeling tasks.

When using a multi-camera setup, 2D tracking results from individual views can be fused by matching geometric features of object detections between cameras [25]. Or, tracking can be performed once in a fused representation of the detection from all views, e.g. an estimated ground plane occupancy map, from per view foreground segmentation [26,27] or object detector responses [28]. Another way to combine multi-view images is to project the segmented foregrounds in different calibrated views to the ground plane, called homography [29–31]. Taking this concept even further is construct a volumetric representation of the 3D scene [32], which helps to deal with occlusions, and provides additional detailed shape information [33]. In this

Table 1
The features extracted from each harmonic complex.

Feature no.	Description
1	Length in seconds
2	Score from Eq. (16)
3	Feature 2 divided by feature 1
4	Number of signal components
5	Mean energy under the signal components
6	Std deviation of energy under the signal
7	Spectral tilt of the signal components
8	Mean fundamental frequency
9	Standard deviation of fundamental frequency

paper we will take the last approach and use volumetric reconstruction to improve robustness of tracking multiple persons under real-world conditions and occlusions, and to obtain foreground masks associated with (possibly) partially occluded individuals. Within a person's foreground, we compute optical flow as an appearance invariant feature for energetic body movements, and that is indicative for physical aggression.

2.2. Audio feature extraction

Acoustic aggression detection in minimally controlled open environment requires highly robust sound processing. The acoustical environment changes constantly and multiple sources will be present, many of which were not present during design. While human listeners have no problem dealing with these challenges, and often do not even notice them [34], automatic systems do have grave difficulties. A core problem is that the scope of possible sonic events in possible acoustic environments is much greater than that of any research database. Even more, many sounds resemble verbal aggression, even to human listeners, such as enthusiastic exclamations, and barks of dogs.

Actual verbal aggression is however a rare event, as demonstrated by the verbal aggression detection system that was developed and tested in conjunction with the police of the city of Groningen, the Netherlands [35]. There, each installed detector should classify less than 10 s per month as verbal aggression. The absence of available training data for standard machine learning techniques motivated a knowledge based approach, that forms the basis of the feature extraction approach used here. Evaluation over a 10 weeks period showed that this approach resulted in no false negatives, while false positive events could be reduced from 1359 (with permissive settings to collect more data) to 2 after optimization (on collected samples).

Because verbal aggression is a variant of speech, the detector in [35] is based on a speech detector that was sensitive to speech that shows tell-tale effects of aggression. The influence of aggression is modeled by the component process model from [36], where two emotions closely related by aggression, namely anger and panic, are treated as ergotropic arousal. This form of arousal is accompanied by increase of heart rate, transpiration and associated hormonal activity. For speech this results in more stress on the vocal folds, which in turn results in a higher, more unstable, pitch (features 8 and 9 in Table 1) and a shift of energy to the higher frequencies (feature 7). This response is in line with the Lombard reflex [37], which occurs if a speaker wants to be noticed over competing sources. [35] showed that these generic features are selective enough for aggression detection. Their system performs foreground–background separation in combination with an analysis of the energy distribution, pitch extraction of the foreground signal and pattern matching. The background model is a first order model with a time constant of 10 s that is dynamically updated when the local energy in a cochleogram (a spectrogram derived from an auditory model) is within 6 dB of the current background model value. This leads to a foreground that contains all information that changes rapidly compared to the background. In normal social conditions the foreground is likely to represent multiple sources. Since pitch extraction is based on the whole foreground, the system is sensitive to erroneously interpreting concurrent pitch tracts as a single pitch.

The conceptual improvement we implement in this paper is the use of tonal signal components, which arise as a string of peaks in the cochleogram, in combination with a pitch extraction algorithm that selects and combines harmonically related tonal signal components into harmonic complexes. These are highly likely to contain tonal information of a single source, unlike the foreground selection of [35]. The harmonic complexes are used to determine whether they might be a voice, and if so, whether the voice is sufficiently shifted toward aggression to justify an alarm.

2.3. Fusing multi-modal data streams

Fusing observations from multiple modalities has shown promising results in various applications. For instance, a combined microphone array and camera setup can improve a particle filter for tracking over using a single modality only [38], since the audio and video complement each other in cases where one modality would lose track due to noise or occlusion. In [39] multi-modal speaker diarization (i.e. the problem of determining who is speaking when) enables automated camera panning during conference calls or improved multi-person interaction with robots.

Bayesian Networks (BN) have been used to combine multiple sources of evidence probabilistically, such as tracked object velocity and position in the scene plus local image features [14], and to model their relation to latent variables of interest [40]. In [23] a single BN represents the whole scene for offline analysis, with observed variables for detected atomic events, and hidden variables that link the detections into larger compound events of interest. The Dynamic Bayesian Network (DBN) additionally models the temporal dynamics of a latent processes, and is therefore commonly used in activity recognition tasks [8,9,15,16,19,40–43].

Lefter et al. [40] discuss multi-modal aggression detection within confined public transportation vehicles, where it is possible to extract linguistic features (e.g. detected aggressive keywords such as cursing) in addition to audio and video features. They evaluate several approaches to exploit meta-features that encode when to rely on specific modalities (the DBN was however too complex for their approach, as the meta-features introduce many latent variables). For similar scenarios, Vu et al. [44] propose a declarative knowledge base to construct high-level event descriptions from observed low-level audio and video events. However, appropriate rules that account for temporal integration and detector confidence need to be constructed manually. [42] tailors various knowledge representation frameworks, such as rule-based reasoning and Bayesian modeling, to detecting aggression within train compartments, and presents some qualitative experiments on human-annotated data.

In surveillance scenarios where it is unknown or hard to define what constitutes undesired or anomalous behavior, an alternative is to create a model of the normative data only, and flag anything out of the ordinary. Such anomaly detection therefore involves unsupervised learning, such as data clustering with outlier detection, or density estimation with a likelihood threshold [43,45–47]. The multi-modal violence detection system proposed by [43] targets fights in urban environments, and utilizes thermal imaging and a microphone array in addition to video data. Sensor fusion and temporal integration are achieved by combining event streams from the individual sensors in a single hidden Markov model trained on normative behavior. For test sequences, the likelihood under this model is computed at each time instance, and instances where it is lower than a given threshold are considered anomalies. They show that anomalies are more prevalent during fights, but are also incurred by moving vehicles in the background. [41] classifies event sequences in video

as normal or anomalous in an unsupervised manner, but uses generative models for both classes (namely, a mixture of DBNs). A new input sequence is classified as normal only if its passes a likelihood ratio test. Then, the sequence is used to update the model parameters for the assigned class, such that the model adapts online to common use cases without supervision.

Cristani et al. [48] propose to automatically discern events independently in audio and video first, and then fuse these in an Audio–Video Concurrence (AVC) matrix to encode the degree of co-occurrence between the events in both modalities. The AVC matrix can be used to segment the input streams online, and also as a feature for event classification. Here it is assumed that simultaneously occurring A/V events are likely to be causally correlated (e.g. a person appears when a phone rings). The Audio-Visual Grouplets presented by [49] are a bag-of-words representation for the foreground and background of both audio and video data. The bag-of-words representation is designed to discriminate between different classes of generic video sequences, such as 'wedding' or 'basketball', and therefore attempts to blindly separate fore- and background in both modalities without exploiting scene specific knowledge.

Finally, there is also research on multi-modal violent scene detection in TV series and movies. In [50–52] detect genre specific events in both audio and video (e.g. flames and explosions, gun shots and bloody imagery). Unlike in the surveillance scenarios, temporal dynamics are not modeled, and no attempt is made on the video side to detect and track people. Instead, classifiers are trained to label individual scenes as either violent or non-violent based on features.

In our approach, the motion features from different views are first fused per person at the feature level, taking into account a person's projected size in each view. Further, we use a DBN to fuse contextual information with observations from both video and audio sources, and model the temporal relation between the latent aggression state at each time instance. While we do not intend to describe interperson relationships at the level of social relationships, we do include person interaction based on proximity in our model to assess the aggression threat. Additionally, we take into account the proximity between observed aggressive behavior and certain static objects in the scene, as it may be indicative of vandalism. The need for such spatial context to interpret activity in surveillance video has been noted previously by [9].

3. System overview

The proposed system addresses the automatic detection of aggressive human behavior in public environments, such as a train station, with non-scripted activity in the background (people passing by, trains stopping and leaving), changing illumination conditions (e.g. shadows), and uncontrolled audio noise. The system uses one microphone and three calibrated overlapping cameras. It conceptually consists of a video, an audio, and a sensor fusion unit. The *video unit* (Section 4) tracks individuals and extracts visual aggression features from the motion field of the most energetic person in the scene, and the distances to the closest nearby person or marked object of interest. The *audio unit* (Section 5) identifies specific sound events in the input signal, using detectors trained for a selected set of audio classes. At each time step, the features/events are combined in the *fusion unit* (Section 6), which introduces temporal coherence and yields as output of the CASSANDRA system an estimated overall aggression level.

The system is an improvement over our earlier work [53], where a single camera view was used and only one audio class was detected. In that system, persons were tracked as ellipsoid regions in the image plane; it was therefore more sensitive to inter-person occlusion, motion in the background and scaling of optical flow features due to perspective. Furthermore, no measure of interaction was included. Section 7 provides an experimental comparison with [53], and also additional experiments on more data.

4. Video unit

The first task of the video unit is to track multiple persons in the 3D scene (Section 4.1), which involves suppressing false positive detections (Section 4.2), and solving a data association problem (Section 4.3). For the tracked persons a pseudo-kinetic energy measurement is obtained (Section 4.4) as a visual aggression feature. Additionally, a person interaction feature based on person proximity is computed as a second visual cue. The video unit also includes a train detector (Section 4.5) for contextual information to reduce the influence of environmental noise in the aggression assessment.

4.1. Multi-person tracking

As discussed in Section 2, there are various methods for multi-person tracking from overlapping views. We build here upon our earlier work [33], and compute at each time step a binary foreground mask for each camera, using the method of Zivkovic and van der Heijden [54]. Since foreground regions are noisy and may merge occluding individuals, the masks are combined into a volumetric representation through volume carving [55]. This representation will be used for tracking, and to label each person in the foreground mask while accounting for occlusion.

Volume carving [32], illustrated in Fig. 2, divides the spatial volume into small 3D grid cells called voxels (analogous to pixels in a 2D image). Each voxel has a binary state, identifying whether it 'remains' or is 'removed'. Initially, all voxels remain, but subsequent carving steps with each camera's foreground mask will incrementally remove more voxels, see Fig. 2b–d. Since each camera is calibrated, it is possible to project the voxel positions to pixel coordinates. Thus, given a binary foreground mask, any voxel that projects to a background pixel does not explain the observed foreground, and must be removed. After carving with all cameras, the remaining voxels are those that correspond to foreground in all camera masks simultaneously.

The obtained voxel regions represent possible body mass of persons in the scene, Regions that are significantly larger than a single person of average size may represent multiple persons or may be caused by segmentation errors. An expectation maximization (EM) based method [33] is applied to locate candidate object volumes within the carved volume, under the constraint that sufficient voxel mass must be present at each found position to contain a human body (see circles in Fig. 2d). The number of candidates is estimated by dividing the region mass by a person's average size, and too small regions are discarded directly. Then, the voxels that constitute an object's volume are labeled with the object's id. Due to incorrect correspondences between the foreground segments across views, carving typically retains more voxels than necessary. As a result, additional candidate objects will be found, which we term *ghosts* as these do not correspond to any actual person in the scene. The number of possible mismatches increases exponentially with the number of objects in the scene. The next section therefore describes a scheme to identify and discard such ghosts. The final voxel representation can generate labeled foreground masks for the remaining objects, accounting for occlusion, in any (camera) viewpoint (Fig. 2e). This representation will then use for data association (Section 4.3) to assign non-ghost objects to tracks.

4.2. Ghost detection

Ghosts are false positives within the set of objects detections $\mathbf{O}$ found in the carved voxel volume. In order to remove such ghost objects, we introduce a probabilistic formulation to identify a minimal subset of objects whose labeled voxels sufficiently explain the observed foreground. Note that ghosts project to less foreground than their non-ghost counterparts (approximately, due to segmentation

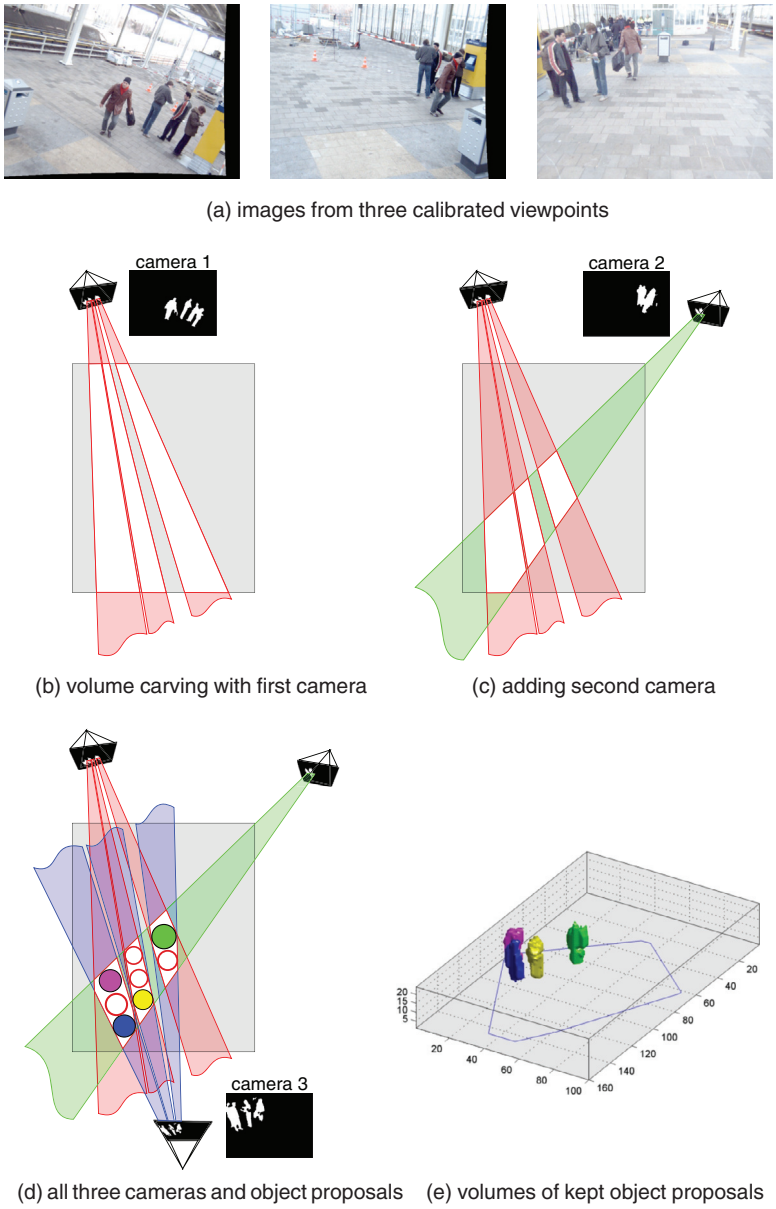

Fig. 2. (a) Volume carving uses multiple overlapping camera views, for which binary foreground masks are computed. (b) 3D volume, shown top-down as 2D rectangular area, carved with foreground mask of first view. In practice, the volume is represented as a dense grid of voxels. Only volumetric locations (i.e. voxels) that project to the foreground (following the red projection lines) remain (shown in white), other locations are removed (shown in gray). (c) Additional views used for carving will only remove more volume. (d) After carving with the last view, an EM based method locates candidate objects (shown as circles) within the remaining volume. The data association step determines which objects correspond to tracks (filled circles, pseudo-color represents track id), and which are 'ghosts' (red circles). (e) Voxels assigned to tracked objects (shown in side view with track pseudo-colors) provide occlusion aware foreground masks for any viewpoint. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

noise), as each volume carving step with another view discards more voxels where foregrounds mismatch.

For any given time instance, we use the notation F_c to denote the foreground region (i.e. the pixels contained in the foreground segments) in camera $c \in C$, where C is the set of all cameras, and $\mathbf{F}$ is the vector of all foreground regions. For a given object $o_j \in \mathbf{O}$, we define $\text{proj}_c(o_j^v)$ as the image region obtained by projecting the voxels o_j^v onto camera view $c \in C$. Similarly, for a subset of objects $O \subseteq \mathbf{O}$, we define $\text{proj}_c(O)$ as the total image region in camera c of all objects, i.e. $\text{proj}_c(O) = \bigcup_{o_j \in O} \text{proj}_c(o_j^v)$. The function $\text{overlap}(\text{proj}_c(O), F_c)$ describes the fraction (within range [0, 1]) of the segmented foreground that intersects with the projected image region. Note that any projected voxel region in camera c is always contained in the foreground region F_c, thus $\text{proj}_c(o_j^v) \subseteq F_c$ for all j, c. Therefore, $\text{proj}_c(o_j^v)$ can also be interpreted as the common intersection-over-union measure for binary regions, and is efficiently computed as the number of pixels in the projection divided by the number of pixels in the foreground.

The probability $P(\mathbf{F}|O)$ of the observed foreground in all cameras given a set of objects $O \subseteq \mathbf{O}$ is modeled using the overlap between the observed foreground regions and predicted regions, i.e.:

$$P(\mathbf{F}|O) \propto \sum_{c \in C} \text{overlap}(\text{proj}_c(O), F_c). \tag{1}$$

The a-priori probability that a subset contains only non-ghost objects decreases as the subset size increases. The optimal set of objects $O^\star \subseteq \mathbf{O}$ for given foreground segments $\mathbf{F}$ is therefore found as the

following maximum a-posteriori (MAP) estimate,

$$O^\star = \arg\max_{O \subseteq \mathbf{O}} P(O|\mathbf{F}) = \arg\max_{O \subseteq \mathbf{O}} \big[P(\mathbf{F}|O)P(O)\big]. \quad (2)$$

To determine subset $O^\star$ one could try to evaluate all possible subsets exhaustively, but this quickly becomes intractable as the number of objects increases. Instead, all sets of size n are evaluated before larger ones of size $n+1$, finding $O^\star$ in a breadth-first fashion. Given a set O of size n, we observe from (2) that a set $O^\dagger = O \cup \{o_j\}$, with $o_j \notin O$, will only have a higher probability than O if and only if

$$\frac{P(\mathbf{F}|O^\dagger)}{P(\mathbf{F}|O)} > \frac{P(O)}{P(O^\dagger)}. \quad (3)$$

In our model, the prior $P(O)$ only depends on the number of objects $|O|$ such that $\frac{P(O)}{P(O^\dagger)} = \eta$, where η is a constant. Thus to add o_j to the set of real objects, the ratio at the left-hand side of Eq. (3) should exceed this constant (user defined in the experiments). Otherwise, the addition of o_j to O yields a suboptimal solution, and any set $O\dagger$, with $O \subset O\dagger$ and $o_j \in O\dagger$, can be pruned from future evaluation.

4.3. Data association

The data association problem in the video unit involves assigning detected objects to the available tracks at the current time step. Each tracker $t_i \in \mathbf{T}$, from the set of trackers $\mathbf{T}$ of the previous time step, has a ground plane position t_i^l and appearance estimate t_i^a. Track assignment can be seen as an edge selection task on a bipartite graph, where one set of nodes represent the existing tracks and the other set of nodes represent the segmented objects. Assuming for the moment, that no tracks are added or deleted, assignment A is a set of (o_j, t_i) pairs, such that all $o_j \in O^\star$ and all $t_i \in \mathbf{T}$ occur exactly once. We are interested in the assignment which maximizes

$$P(A) = \prod_{(o_j,t_i)\in A} P(o_j, t_i) \propto \prod_{(o_j,t_i)\in A} P^{\text{loc}}(o_j^l|t_i^l) P^{\text{app}}(o_j^a|t_i^a) \quad (4)$$

where $P^{\text{loc}}(o_j^l|t_i^l)$ and $P^{\text{app}}(o_j^a|t_i^a)$ are defined below. The above combinatorial problem can be solved efficiently with the Hungarian algorithm [56].

The location likelihood is defined as

$$P^{\text{loc}}(o_j^l|t_i^l) \propto e^{-\lambda D^E(o_j^l, t_i^l)} \quad (5)$$

where $D^E(o_j^l, t_i^l)$ the Euclidean distance between the location of the detected object o_j^l and of the tracker t_i^l.

Object appearances are represented as three 3D color histograms (R, G and B channels): one histogram for the legs, arms/torso and head/shoulders region, respectively. Occlusion order and visibility is taken into account by measuring within the person masks given by $O^\star$. This simple part-based representation allows to deal better with inter-person occlusion. Histograms are extracted from each camera viewpoint and subsequently averaged. The appearance likelihood is defined as

$$P^{\text{app}}(o_j^a|t_i^a) \propto e^{-\kappa D^B(o_j^a, t_i^a)} \quad (6)$$

where $D^B(o_j^a, t_i^a)$ is the Bhattacharyya distance between the histograms of the object and the filtered histogram estimate of the tracker. Wen computing the color histograms.

To allow for track creation and termination, extra nodes are added to the bipartite assignment graph mentioned earlier. Assigning a tracker to one of these nodes discontinues the track, while the assignment of a segmented object to one of the extra nodes creates a new tracker. To determine the likelihood of new or discontinued tracks, the appearance term on the right hand side of Eq. (4) is replaced by a constant factor, and a location likelihood that is determined by a spatial map which encodes that track creation and termination is more likely to occur near the edge of the scene [33].

Finally after object to tracker correspondences have been made, the measured position filtered by means of a Kalman filter, and appearance histogram bins are updated by an exponential decay function, i.e. $t_i^a \leftarrow (1-\alpha)\cdot t_i^a + \alpha \cdot o_j^a$.

4.4. Video feature extraction

The voxel-based person tracker provides per frame the locations of people in the scene, and their non-occluded image region in each camera view. Within these image regions visual features are extracted that are indicative of body articulation. We describe the human body as a collection of points with identical mass. While such a model is clearly a simplification, it reflects the non-rigid nature of a body well and facilitates fast computations. In each camera c we select 100 points within the visible image area of a tracked person j by finding pixels with the most local contrast [57]. Such points are easy to track and usually align well with edges in an image (which in turn often coincide with limbs, as seen in Fig. 3 top, bottom left). The KLT algorithm [57] is used to track points within subsequent images, resulting in 100 displacement vectors in image coordinates. Outliers, displacement vectors for which the length of the vector is larger than twice the standard deviation of all vector lengths are discarded. Also vectors which are classified to be part of a passing train are discarded (see Section 4.5). In total, we are left with Q relevant displacement vectors that will be used measure the amount of kinetic energy in a person's movements.

Two operations are performed on these displacement vectors. First, in order to discount overall body motion (e.g. as induced by walking) and only capture the relative articulation energy of the limbs, the mean displacement is subtracted from all displacement vectors in a single view. Note that displacement vectors are measured over all visible body parts, thus not only on moving limbs. For instance, the mean displacement of a static person raising an arm is near zero and does not cancel the arm motion. Second, to correct for the perspective projection, the magnitudes of the displacement vectors in each view are scaled by the distance of the person to the camera. An advantage of the calibrated camera setup is that this is straightforward, as the relative position in meters of the (tracked) person to each camera is known. We thus obtain from all views perspective-invariant velocity vectors v_q, for $q = 1 \ldots Q$, with which the pseudo-kinetic [53] energy $\bar{E}_j$ of person j is computed,

$$\bar{E}_j = \frac{1}{Q}\sum_{q=1}^{Q} |v_q|^2. \quad (7)$$

This feature provides a single measure for the intensity of a person's articulated motion, and will be our primary visual cue for aggression detection.

In a scene with multiple persons, we assume that the most energetic person is indicative for the ambient scene aggression level. The ambient pseudo-kinetic energy output feature ϕ_k at time step k is thus computed as

$$j_{\max} = \arg\max_j \bar{E}_j \quad (8)$$

$$\phi_k = \max_j \bar{E}_j = \bar{E}_{j_{\max}}. \quad (9)$$

Furthermore, the interaction of this person with other people, or objects in the environment, is indicative for aggression too. A detailed understanding of person interaction would require a high-level semantic interpretation of the scene, which is currently out of the scope of the CASSANDRA system. Instead, we take proximity of the most energetic person to the nearest person or object as a proxy for interaction, which is sufficient for our needs. Intuitively, strong limb

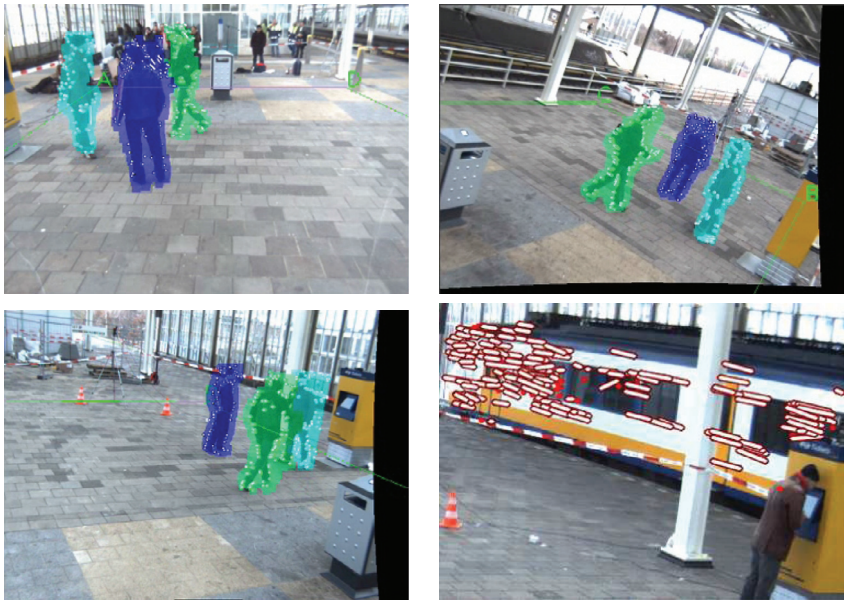

Fig. 3. Top, and bottom left: tracking results for a selected frame from the three camera perspectives. The voxels are rendered with a different color, depending on the tracker. On top of the images the displacement features are shown that are used to compute the kinetic energy per tracker. Bottom right: optical-flow features for detecting trains in motion.

movement is considered more aggressive when one is standing close to another person or object, whereas when one is standing apart from others it is indicative of harmless waving, stretching, etc.

The interaction measurement ξ_k at time step k is then computed as the minimum Euclidean distance between the tracked position of the most energetic person $j_{\max}$ (Eq (8)) and the set $L_{j_{\max}}$ containing the locations of all other tracked persons and the physical objects, thus

$$L_{j_{\max}} = \{t_j^l \mid \forall t_j \in \mathbf{T},\ j \neq j_{\max}\} \cup L' \tag{10}$$

$$\xi_k = \min_{l \in L_{j_{\max}}} \left[D^E(l, t^l_{j_{\max}})\right]. \tag{11}$$

Here the set L' contains predefined locations of physical objects of interest in the scene (e.g. a ticket vending machine), such that acts of vandalism by even a single person are detectable as a form of aggressive interaction.

4.5. Train detection

An additional objective of the video unit is to detect moving trains. Trains moving in and out of a station produce visual and auditory noise that may lead to spurious aggression detections. Therefore, recognizing trains in video opens a possibility for suppressing such noise both at the signal level and later in the fusion unit. A train 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 $N = 100$ KLT motion features [57] are tracked frame-to-frame (Fig. 3, bottom right). The motion vectors are classified as train/non-train by testing if size and direction are within preset bounds. The state of the train detector for that region, which is fed to the fusion unit, is active when more than 50% of the features are classified positively. Due to the constrained movement of trains, our simple detector turns out quite robust to occasional occlusions of the train area by people. A person's foreground mask could accidentally include motion from a train in the background. Therefore we filter articulation features found in a region where a train in motion is detected to prune misdetections. Since the observed length of train flow features depends on the train velocity and distance to the camera (due to perspective), only articulation features that are sufficiently similar to the nearest observed train flow features are removed. Body part articulation and train motion only coincide sporadically, thus true articulation features are rarely affected.

5. Audio unit

For our audio detections, we separate speech, singing, kicking-object and screaming from other sounds. Here, 'singing' is used as the class label for various cases of chanting supporters in the dataset, and can be a precursor for aggression later in the fusion unit. To do this in a robust way we focus on the tonal components in the signal. Tones are the basis of voiced speech and are robust to inference of other sources because they are sparse in frequency and therefore overlap little with other sources. Moreover all energy is concentrated in one frequency therefore they are likely to have a positive local signal-to-noise ratio. This approach is described in [58].

We extract tones in the time–frequency domain. To convert the audio signal to the time–frequency domain a gamma-chirp filterbank [59] is used. The filterbank consists of 100 channels with filter-coefficients $h(t)$ following

$$h(t) = at^{N-1}e^{-2\pi bB(f_c)t}e^{j(2\pi f_c t + c\log(t))}, \tag{12}$$

where f_c is the center frequency of the channel, N the order of the gamma-chirp ($N = 4$) and $a = 1$, $b = 0.71$ and $c = -3.7$. The center-frequencies are logarithmically distributed between 60 and 400 Hz. B is the bandwidth of a filter and is given by the ERB scale [60]:

$$B(f_c) = 24.7 + 0.108 f_c \tag{13}$$

The choice for a gamma-tone filterbank is based on its good time–frequency localization [61], while keeping reasonable noise-robustness [62]. The filter output A_n is squared, leaky-integrated with channel-dependent time-constants ($\tau_c = \frac{2}{f_c}$, Eq. (14)) and finally down-sampled to 200 Hz (Eq. (15), where s is the sampled frame number, and $\Delta t_s = 5$ms). Taking the logarithm results in a log-energy representation called a cochleogram, an example can be seen in Fig. 5.

$$E_n(t) = \int_{t_0}^{t} A_n^2(t-\tau)e^{-\tau/\tau_c}d\tau \tag{14}$$

$$E_n^{dB}(s) = 10\log_{10}(E_n(s\,\Delta t_s)). \tag{15}$$

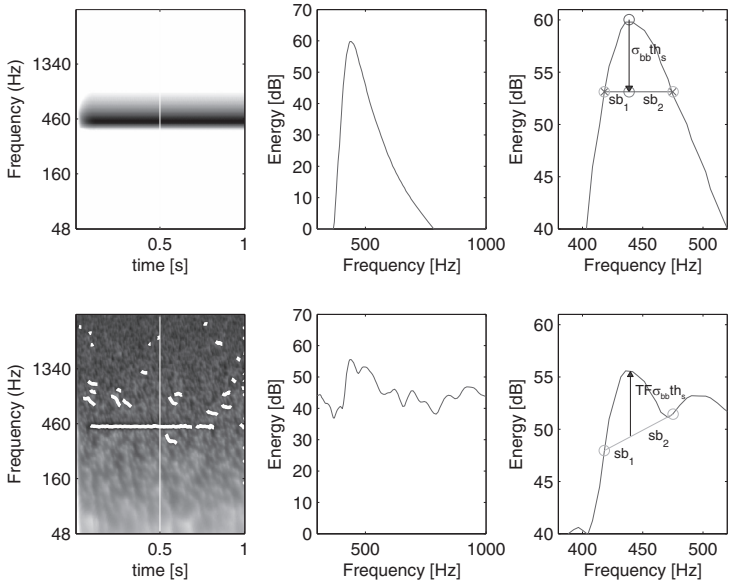

Fig. 4. Computation of the tone-fit (TF). The upper left panel shows an ideal sinusoid, the lower a noisy sinusoid with a decreasing SNR and the extracted signal components, including some spurious ones. The upper middle panel shows an ideal sinusoid response around the peak (in the channel direction). The lower middle panel shows the cross section around the noisy pulse. The upper right panel shows the computation of the TF at the peak position. The TF is the energy difference denoted by the vertical line. The lower right panels shows the TF computation for the noisy tone.

To extract voiced speech from the cochleogram we start by estimating the local tone-likeness of every point in the time–frequency plane. The response of the cochleogram to tones is very predictable and robust to interfering sources up to 6 dB local target-to-non-target ratio. The tone-likeness is measured with a matched filter. This filter has a width in frequency direction of the response of a perfect sinusoid. This width is determined at th_s, twice the standard deviation of the energy of broadband noise (σ_{bb}) under the energy maximum of the tone. Because of the logarithmic frequency axis, the width is asymmetric and therefore two widths are recorded (sb_1 and sb_2). The upper panels of Fig. 4 illustrate this. The normalization by σ_{bb} ensures that the amount of spurious the peaks in broadband signals is frequency independent and predictable. The application of the filter is the reverse process and is illustrated in the lower panels of Fig. 4. The difference between the expected energy, the weighted average of the energy sb_1 and sb_2 (respectively below and above the frequency in question), and the actual energy (normalized by $\sigma_{\text{bb}} th_s$) is the tone-fit measure. This measure is frequency-independent and equals 1 for perfect sinusoids.

The tone-fit is applied to every point in the time–frequency plane, the resulting matrix is thresholded (tone-fit > 0.5) and all connected components in the resulting mask are extracted. Components with an area larger than what can be expected in noise are accepted as tonal-components, the others are discarded. Within the accepted components the energy maxima are strung together to form a sparse representation of the tonal components. Due to the filter properties only a single energy maximum can exist per frame. This relieves the demand for, for example, McAulay–Quatari tracking [63]. These tonal components have a high probability of stemming from a single source. They are depicted as thin white lines in Fig. 5.

Co-developing tonal components are grouped together based on common fate principles [64]. The algorithm generates multiple grouping hypotheses and these are scored according to:

$$S = n_{\text{sc}} + b_{f0} + n_h - \sum_{\text{sc}} \text{rms}_{\text{sc}} - \sum_{\text{sc}} \Delta f_{\text{sc}} \tag{16}$$

where n_{sc} is the number of signal components in the group, b_{f0} is one or zero depending on the existence of a signal component at the fundamental frequency, n_h is the number of sequential harmonics in the group, rms_{sc} are the root mean square values of the difference of a signal component and the fundamental frequency after the mean frequency difference is removed, and Δf_{sc} is the mean difference between the fundamental frequency and the frequency of the signal component divided by its harmonic number. This scoring function is identical to Eq. (3) in [58]. The hypothesis with the highest score is picked and used as the basis of recognition of speech, singing, screaming and outlier.

The features extracted from the harmonic groups are listed in Table 1. The features for discriminating speech, singing, kicking-objects and screams are based on two properties: the strength of the harmonic group (features 1–4) and aggression related properties (features 5–9). The last set of features is based on [35,53], which in turn are based on research of how the human vocal tract changes under the influence of aggression. The feature vectors are classified with a naive-Bayes classifier from the WEKA-toolbox [65], trained in a leave-one-out setup. After classification the results are delivered to the fusion unit.

6. Fusion unit

The fusion unit produces an aggregate aggression indication given the features/events produced independently by the audio and video units. 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 [66]), to estimate the scene ambient aggression level.

6.1. Basic model

We denote the discrete-time index as $k = 1, 2, \ldots,$ and set the time unit increment to 50 ms. At the kth step, $\psi_k^c \in \{0, 1\}$ denotes the output of the audio detector for audio class $c \in \{speech, scream, singing, kicking-object\}$ (Section 5), ϕ_k denotes the ambient pseudo-kinetic energy, and ξ_k the interaction level (Section 4.4). In the presented system there are four non-overlapping rail sections monitored by three cameras (two cameras monitor a single section, one camera monitors two sections). The output of the mth, $m = 1, \ldots, 4$, train detector (Section 4.5) will be denoted as $y_{m,k}^{\text{T}} \in \{0, 1\}$.

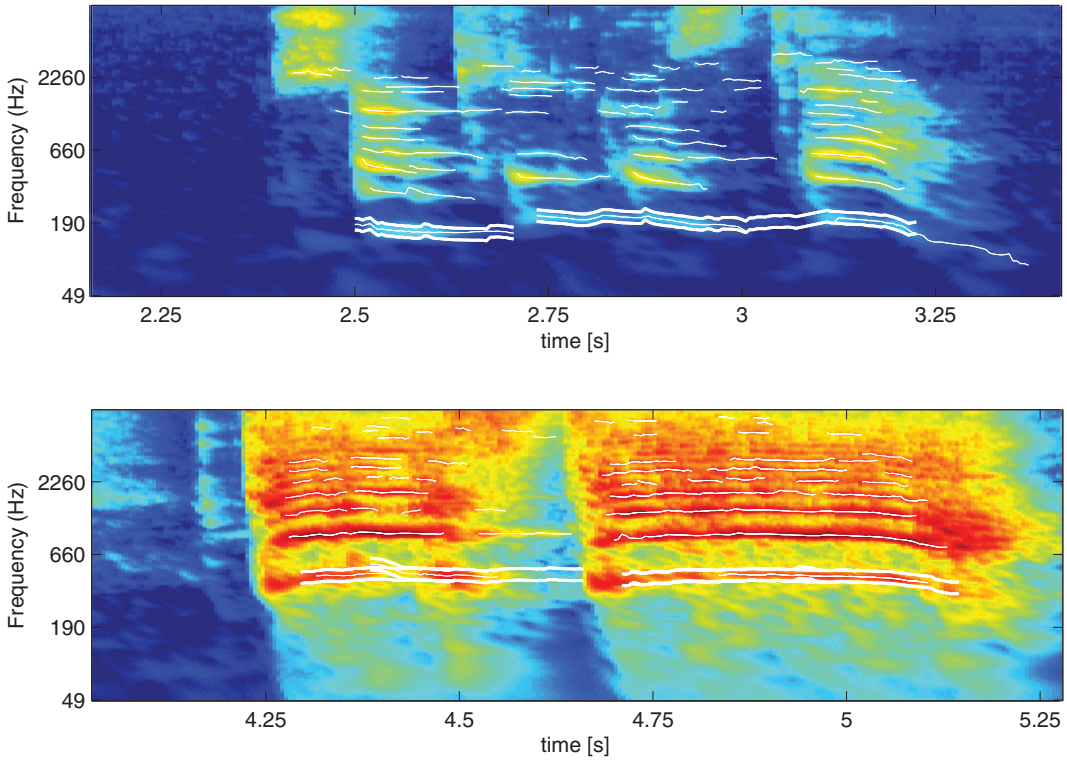

Fig. 5. Cochleograms of aggressive (lower panel) and non-aggressive speech (upper panel). The thin white lines indicate the tonal components as found by the tone-fit algorithm. The double thick white lines indicate the pitch as a result of the harmonic complex extraction.

In order to reason about aggression levels, we use a five step discrete scale $\langle 0, 1\rangle$: 0.0 (no activity), 0.2 (normal activity), 0.4 (attention suggested), 0.6 (minor disturbance), and 0.8 (major disturbance) up to 1.0 (critical aggression). The visual aggression features ϕ_k and ξ_k are discretized into four steps.

The aggression level obeys specific correlations over time and is represented as a process rather than an instantaneous quantity. We denote the aggression level at step k as a_k and define a stochastic process $\{a_k\}$ with dynamics given by a first-order Markov chain with the following state transition probability:

$$p(a_{k+1} = i | a_k = j) = \mathrm{CPT}^a(i, j), \tag{17}$$

where $\mathrm{CPT}^a(i, j)$, denotes a conditional probability table. While this transition formulation does not enforce an ordered relationship between levels, i.e. a_k is categorical rather than ordinal, transitions between neighboring levels will be more probable since this is reflected by the aggression level transitions in the training data.

The measured visual (ϕ_k, ξ_k) and auditory ($\psi_k = \{\psi_k^c\}$) features are treated as samples from an observation distribution that depends on the aggression level a_k. Since 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 also depends on the train-noise indicator:

$$p(\phi_k, \xi_k, \psi_k | a_k, n_k) = p(\phi_k | a_k) p(\xi_k | a_k) \prod_c p(\psi_k^c | a_k, n_k) \tag{18}$$

The model takes the form of conditional probability tables CPT^ϕ and CPT^ξ for the visual aggression features, and CPT^{ψ^c} for the audio class detections ψ^c.

$$p(\phi_k = i | a_k = j) = \mathrm{CPT}^\phi(i, j), \tag{19}$$

$$p(\xi_k = i | a_k = j) = \mathrm{CPT}^\xi(i, j), \tag{20}$$

$$p(\psi_k^c = i | a_k = j, n_k = n) = \mathrm{CPT}^{\psi^c}(i, j, n). \tag{21}$$

6.2. 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 [53]. Additionally, each train model encodes prior information about the 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

$$p(y^{\mathrm{T}}_{m,k} | i_{m,k}) = \mathrm{CPT}^t(y^{\mathrm{T}}_{m,k}, i_{m,k}) \tag{22}$$

$$p(n_k | i_{1:M,k}) = \mathrm{CPT}^n(n_k, i_{1:M,k}). \tag{23}$$

For each rail, Eq. (22) encodes inaccuracies of detector (misdetections, false alarms). Eq. (23) 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 s (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

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

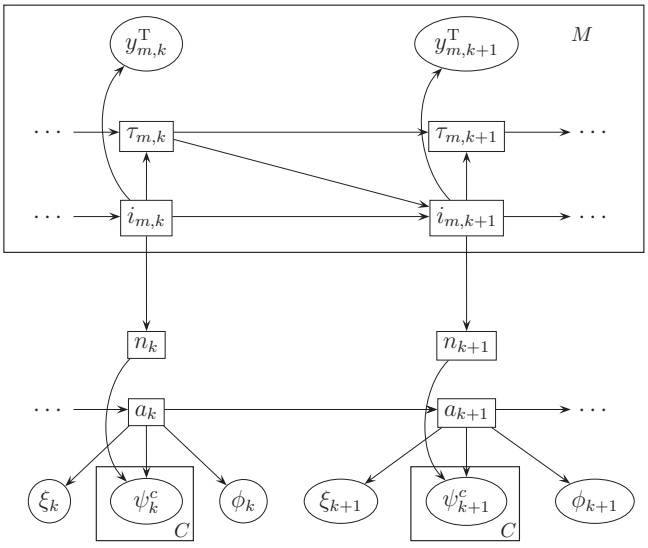

Fig. 6. Dynamic Bayesian Network representing the probabilistic fusion model. The rectangular plates indicate replications for the $C = 4$ audio classes and $M = 4$ train detector sub-networks. Oval nodes denote observed variables, square nodes are hidden.

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 a 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 $\epsilon = 50$ ms is the period between successive steps.

6.3. Inference and parameter estimation

In the 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 up to the current step, $p(a_k|\phi_{1:k}, \xi_{1:k}, \psi_{1:k}, y^{\mathrm{T}}_{1:m,1:k})$. From this distribution we calculate the expected aggression value, which will be the basic output of the fusion unit.

Given the graphical structure of the model (Fig. 6), the required distribution can be efficiently computed using a recursive, forward filtering procedure [66]. We implemented an approximate variant of the filtering procedure, known as the Boyen–Koller algorithm [67]. At a given step k, the algorithm maintains only marginal distributions $p(h_k|\phi_{1:k}, \xi_{1:k}, \psi_{1:k}, y^{\mathrm{T}}_{1:m,1:k})$, 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, the aggression level evolves usually more subtly as the tension and anger among people build up. We additionally enforce temporal smoothness by applying a simple low-pass filter to the (pseudo-)kinetic energy and person interaction measurements (before inference) and the expected aggression level (after inference).

The parameters of probability tables CPT^a, CPT^ϕ, CPT^ξ, CPT^ψ, CPT^n, CPT^t, and the parameters α_m, β_m of the gamma pdf's are set to maximum-likelihood estimates on available training data. These data consist of human annotated values for the scene's aggression level $\{a_k\}$ and the train detector and train noise states $\{i_k\}$ and $\{n_k\}$ (see also Section 7.1), plus corresponding observations from the audio and video unit.

In Figs. 7 and 8 the development of the expected aggression level over time with and without a 15 s low-pass filter are shown, including some images at different moments in the scene. In Fig. 7 the filtered output remains consistently high during acts of vandalism which contain some interruptions. The duration of a high aggression state affects the decision to trigger an alarm. The increasing tension in the scenario shown in Fig. 8 is clearly detected by the system, including the fight at the end as a long period of aggressiveness.

7. Experiments

7.1. Dataset

The CASSANDRA dataset was recorded at a platform of the Amsterdam–Amstel train station, a challenging setting from a sensor point of view. The platform is covered by a glass roof, which, given the intermittently sunny and cloudy conditions on the day of the recordings, resulted in strong lighting changes. The glass roof furthermore caused appreciable sound reverberations. Recordings were performed during the normal hours of platform operation, which meant passing, stopping and accelerating trains and metros on the opposing tracks caused significant audio clutter, as well as changes in the visual background. Visual foreground segmentation was further complicated by the presence of moving people in the background.

The dataset consists of 25 runs (1–2 min each), featuring 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). Table 2 gives an overview of the used scenarios, several of which have multiple takes.

A microphone recorded sound at the scene (16 bits, 44.1 kHz sampling rate) and was located about 2 m from the center of the action and about 2 m from the subway track. Video was captured at 20 Hz and 756 × 560 pixel resolution, using three fully calibrated and frame-synchronized cameras. Camera and audio data were aligned using time stamps. The ground truth for the overall scene aggression level was provided by human operators, using the $\langle 0, 1 \rangle$ scale described in Section 6.1; annotation involved successive short fragments of 5 s each. Also, start and stop times of audio events were annotated for as far as these sounds could be clearly distinguished by an experienced annotator (see Table 3). In 10 scenes in the dataset, the ground plane positions of each actor were annotated at every frame.

7.2. Video unit tracking evaluation

We first evaluate the person tracking component of the video unit. At any time step, an estimated person position is compared to that of the closest person in the ground truth, as long as the deviation in the position does not exceed 0.75 m. A person to which a tracker is assigned in at least 75% of the frames in which the person occurs, is considered a true positives (TP), otherwise it is considered a false negative (FN). Trackers which were created, but which did not correspond to any person in the ground truth, are false positives (FP). The detection rate (DR) is the percentage of all persons at all frames that have been assigned to a tracker. Finally, the identity changes (IDC) is a count of how many times a person in the ground truth was tracked by a different tracker. The results are shown in Table 4.

In most scenarios our system is successful in tracking multiple people in the voxel space, and recovers in those cases that a track is lost. We note that performance is significantly worse for scenario 16–2, where a higher group density, person interactions at close proximity (fighting) and people lying on the floor give rise to tracking errors and cause ID changes.

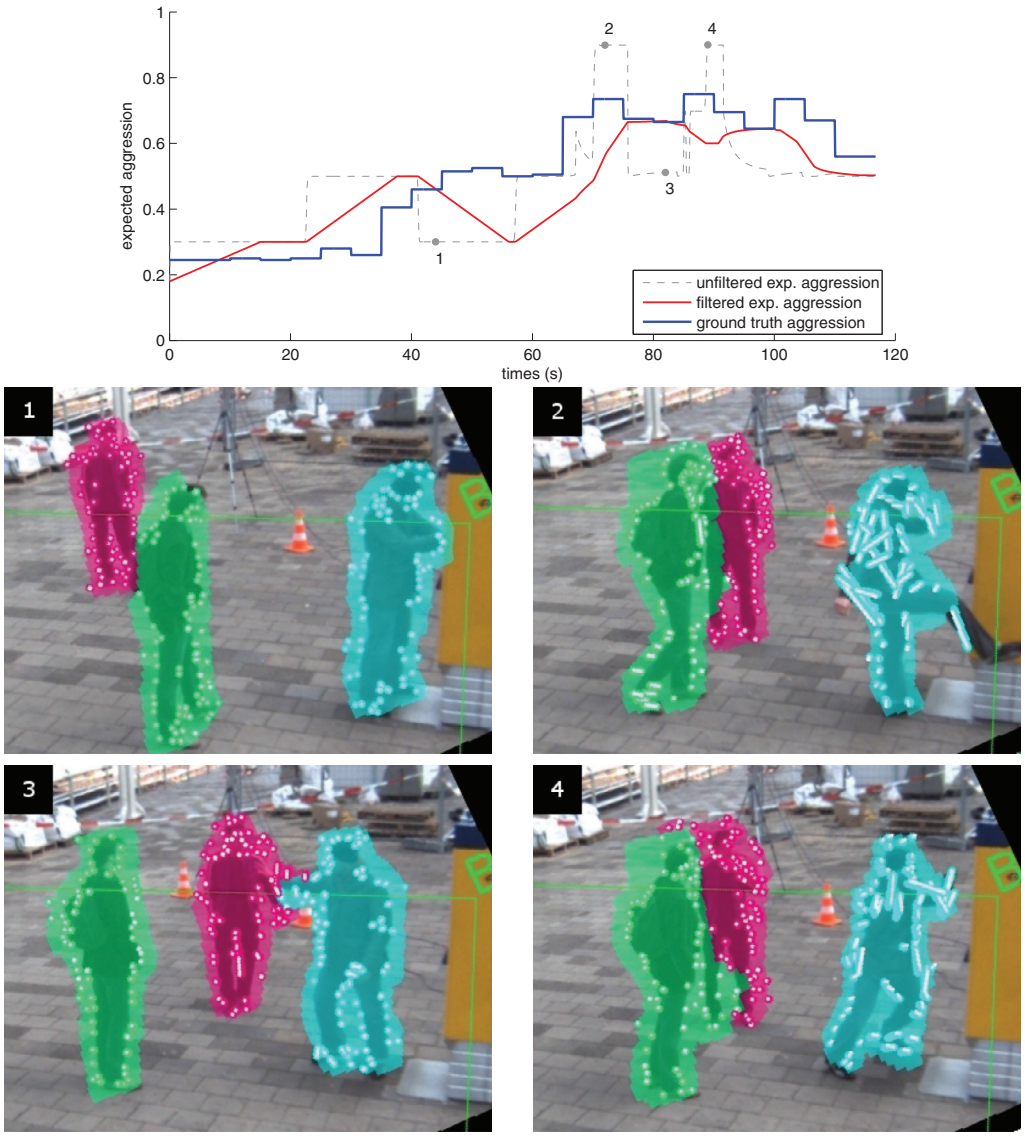

Fig. 7. Aggression built up for a scenario containing a person molesting a ticket vending machine with two bystanders watching in the vicinity. After the beating and kicking the machine for the first time, the aggressor turns to the bystanders, followed by more aggression toward the machine. The top figure shows the development of the expected aggression level over time before smoothing (gray), with the four numbered markers corresponding to the bottom four images. The red line shows the smoothed expected aggression used as system output, the blue depicts the annotated ground truth aggression for the scenario. In the four images the voxel regions and the motion features of the detected persons are colored for visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7.3. Full system evaluation

We evaluated the system with different configurations by including or excluding features from the fusion network:

- `Au`, audio features only;
- `Ke`, kinetic energy feature only;
- `KeAu`, kinetic energy and audio features;
- `KePi`, video features only (kinetic energy and person interaction);
- `KePiAu`, all audio and video features.

A leave-one-out strategy was applied to test the system performance on each of the 25 scenes independently. We considered three quantitative criteria to evaluate the different configurations of the CASSANDRA system: *aggression level error*, *frame based classification*, and *event based classification*.

7.3.1. Aggression level error

This evaluation criterion considers the deviation between the CASSANDRA estimated aggression level and the ground truth annotated aggression level. We compute the mean, standard deviation of the error, plus the root mean squared error (RMSE), see Table 5.

The table shows that as more features are combined the deviation from the ground truth aggression annotation decreases.

7.3.2. Frame based classification

This second evaluation criterion considers aggression detection as a two-class classification problem of distinguishing between "normal" and "aggressive" time steps (ground truth class is obtained by thresholding the ground truth aggression level at 0.5). Similarly, the predicted aggression level can be thresholded, and the resulting classification is compared to the ground truth classification. A trade-off between the true positive and false positive rate is obtained by varying the threshold on the estimated aggression class. Fig. 9 depicts the receiver operating characteristic (ROC) for the different system configurations.

Fig. 9 shows that, using this evaluation criterion, the `KePi` and `KePiAu` configurations have clearly the better overall performance than the other configurations using less features.

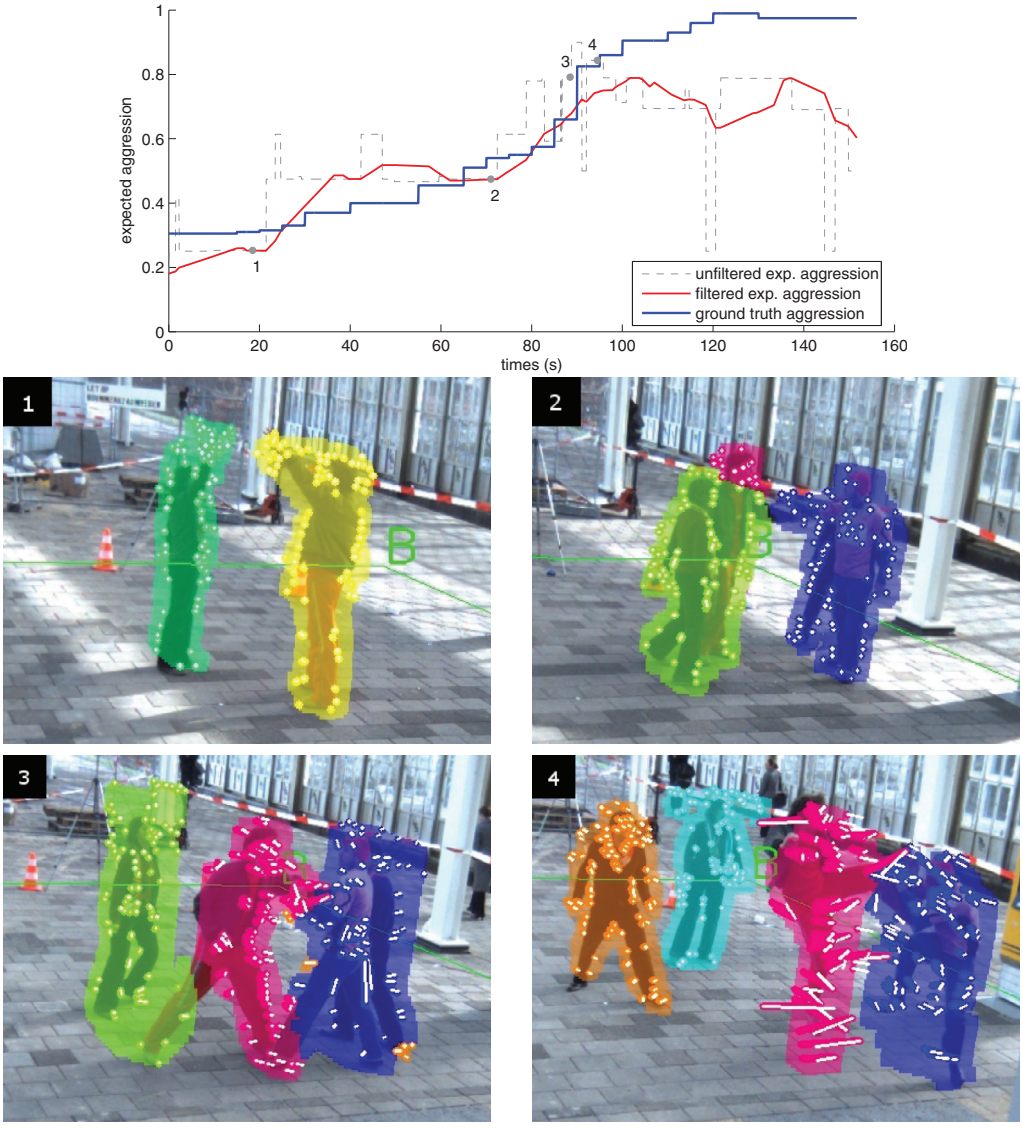

Fig. 8. A scenario containing four supports fighting (two versus two). At the beginning there are only two supports on the platform, but a few moments later two rival support show up, both groups intimidate each other. After that they start fighting (pushing, hitting, kicking). The top figure shows the development of the expected aggression level over time before smoothing (gray), with the four numbered markers corresponding to the bottom four images. The red line shows the smoothed expected aggression used as system output, the blue depicts the annotated ground truth aggression for the scenario. In the four images the voxel regions and the motion features of the detected persons are colored for visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
A summary of the 25 used scenes.

Scenario–take(s)	Description	People
01–1,2	Relaxed; two people meet and hug	3
02–1,2	Lively; two people argue	2
03–1,2	Two people meet and hug	2
04–1	Lively; two people argue	2
05–1	Lively; two people argue, a third person intervenes	3
07–1,2	Normal; various people use a vending machine	4
09–1	Team of thieves; two thieves rob a man	4
10–1,2	Aggression toward a machine; argument	2–3
11–1,2,3	Harassment; a person harasses a passenger	2
12–1,2	Happy supporters; people shouting and dancing	4
13–1,2	Two supporters harassing a third passenger	3
14–1,2,3	Two people fight; a third a person intervenes	3
16–1,2	Supporters fight, two versus two	4

Table 3
The annotated audio classes and the number of their occurrences in the 25 scenes.

Singing	Speech	Kicking-object	Scream
61	386	25	211

7.3.3. *Event based classification*

Instead of treating each frame as an independent classification problem, one could also consider classifying larger time periods as either "normal" or "aggressive" events. The task would then be to identify the highly aggressive events shortly after they have started, while minimizing false alarms. Notice that this task is similar to what the task of a human security officer would be.

An aggression event is a time period during which the aggression level does not drop under a fixed threshold. For this evaluation, aggression events with a total duration less than 9 s are discarded. Every scenario has therefore zero or more aggressive events in the annotated ground truth. Note that various takes of the same scenario are judges individually on observed behavior, and may therefore have somewhat different aggression level annotations. Hence, ambiguous scenarios can have an aggressive event in one take, but none in another take.

Table 4
Tracking results in the video unit for several annotated scenes. people: total number of persons in the scene; TP: true positives (persons tracked for ≥75% time); FN: false negatives (persons tracked for <75% time); FP: false positives (tracker without person); DR: detection rate; IDC: identity changes of all persons; FR: total number of frames in the scene.

Scenario–take	People	TP	FN	FP	DR	IDC	FR
01–1	3	3	0	0	97.9	0	825
01–2	3	2	1	0	88.5	1	892
02–2	2	2	0	0	98.7	0	1013
04–1	2	2	0	0	95.8	0	828
05–1	3	3	0	0	99.0	0	1346
07–1	4	4	0	0	90.7	3	2043
09–1	4	4	0	1	94.8	2	1234
10–2	3	3	0	0	98.8	0	2419
11–1	2	2	0	0	97.3	1	1062
16–2	4	4	0	8	79.6	20	2186

Table 5
Deviation between estimated and ground truth aggression level, for different configurations.

Configuration	Mean	Std. dev.	RMSE
`Au`	0.162	0.154	0.223
`Ke`	0.201	0.145	0.248
`KeAu`	0.152	0.138	0.206
`KePi`	0.162	0.131	0.208
`KePiAu`	0.138	0.128	0.188

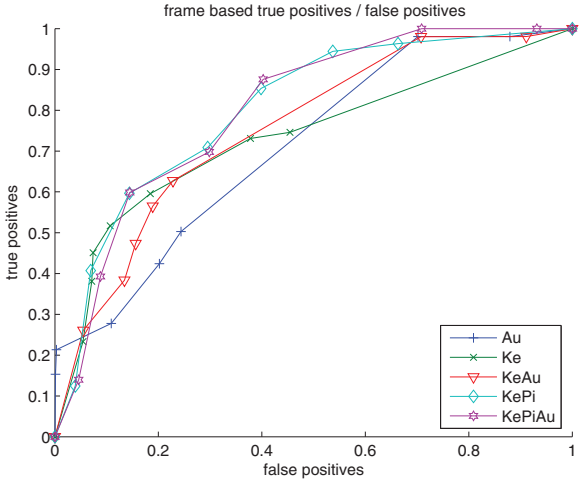

Fig. 9. Frame based classification ROC curve for different CASSANDRA configurations.

Similarly, zero or more aggression events can be predicted by the system. Detected events are compared to the events in the ground truth, considering detections correct when they overlap with the ground truth events. Ground truth events are deemed correctly detected if there is a detection within 10 s after the event started. Detected events which do not correspond to any ground truth event are considered false alarms.

In the 25 scenes used for evaluation, the ground truth annotations contained in total 13 aggression events. Table 6 contains the results for the different system configurations. Notice that the configuration using all available features detects almost all aggressive events on time, while five false alarms are raised over all 25 scenes.

7.4. Comparison with previous work

We compared the results of the system presented in this paper with the results published in [53]. There, aggression was estimated based on audio detector for a single class of "verbal aggression", and motion features in a video stream from a single camera without any

Table 6
Event based classification results for different configurations of the fusion network over 25 scenes.

Configuration	Ground truth events	True positives	False positives
`Au`	13	7	2
`Ke`	13	4	5
`KeAu`	13	6	4
`KePi`	13	10	5
`KePiAu`	13	11	5

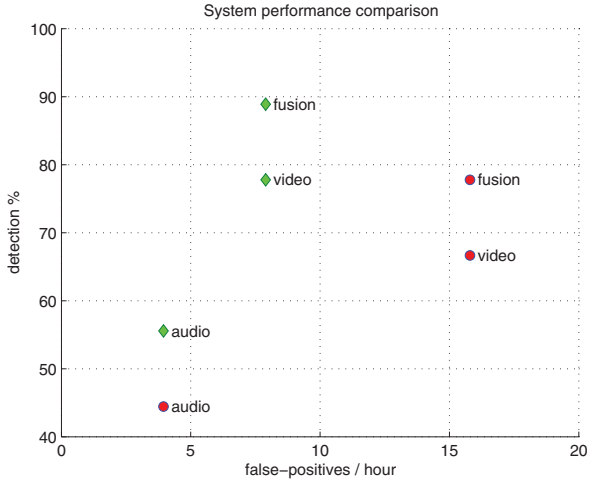

Fig. 10. Comparison of [53] (marked by red circles) with the presented system (marked by green diamonds) on the same 13 scenes, using only audio, or only video features, or all features fused. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3D localization or interaction. The results in [53] were obtained from a subset of only 13 scenes of the scenes used in our evaluation above. For comparison, both systems were trained and evaluated on those same 13 scenes, using a leave-one-out strategy to test each scenario. Again, different configurations were tested using only audio features, only video features, or all features combined. The results shown in Fig. 10, which corresponds directly to Fig. 7 in [53], depict the detection rate of true aggression events versus the number of false alarms raised per hour.

These results show that the presented system improves the overall performance for each sensor modality individually, and that this translates to an improvement of the results when fusing the modalities. A more detailed look at the performance by scenario of both systems is shown in Table 7, which also shows that number of true events in each recording. Recall from Section 7.3.3 that the observed behavior in different takes of the same scenario can be more slightly more alarming in one take than the other, resulting in a different number of ground truth and/or detected events. The table reveals that the presented system still has some difficulties with such ambiguous scenes (e.g. happy football supporters, containing loud singing people, fast movements), though overall less errors are made compared to the reference system. Detection of vandalism against stationary objects has improved, due to the introduction of the interaction cue (Eq. (11)). Apart from ambiguous behavior, we expect that in practice false alarms can occur in atypical situations with respect to the training data (e.g. crowds yield other proximity features), or accidental co-occurrences of various non-critical cues (e.g. shouts from outside the scene while people are gesticulating).

8. Discussion

Our experimental evaluation indicates that there is a clear benefit in the use of complementary video and audio cues, as well as

Table 7
Comparison by scenario of the detections results of the [53] system performance and the presented CASSANDRA (Cas.) system performance. The columns show the number of ground truth positives (GT), plus the true positives (TP) and false positives (FP) per system. Errors are marked in bold.

Scenario description	GT	[53] System		Cas. system	
		TP	FP	TP	FP
Normal: walking, greeting	0	0	0	0	0
Normal: walking, greeting	0	0	**1**	0	0
Excited: lively argument	0	0	0	0	**1**
Excited: lively argument	1	1	0	1	0
Aggression toward a vend. machine	1	**0**	**1**	**0**	0
Aggression toward a vend. machine	1	**0**	0	1	0
Happy football supporters	1	1	0	1	0
Happy football supporters	0	0	**1**	0	**1**
Supporters harassing a passenger	1	1	0	1	0
Supporters harassing a passenger	1	1	0	1	0
Two people fight, third intervenes	1	1	0	1	0
Four people fighting	1	1	0	1	0
Four people fighting	1	1	**1**	1	0

contextual (interactivity, other objects) information, for the estimation of aggression. We found the addition of audio cues to be useful in those scenarios, where there is a build up in aggression, and voices are raised prior to physical assault. During the enactment of physical assault, we found the visual cues to be clearly superior. Compared to the reference system [53], our system also benefits from the improved features obtained using multi-view camera observations and more specific audio classes (see Fig. 10 and Table 7).

The CASSANDRA system runs on two PCs (3.0 GHz Intel processor with 3 GB RAM) one with the audio unit, and one with the video and fusion units. The overall processing rate is on average about 4 s per frame, using un-optimized C and MATLAB code. The processing bottleneck, costing an average of 3 s per frame, is the person segmentation step (Section 4.1). We expect that software optimization and hardware implementation (e.g. DSP, FPGA) will allow real-time processing.

One should be careful not to underestimate CASSANDRA real-world performance, based on Fig. 10. Having 7–8 false positives per hour, per camera, would certainly be too much. It is important to note, however, that our dataset is about equally divided into normal and aggressive time periods. We expect CASSANDRA to produce much less false alarms per hour in a typical surveillance setting, where (fortunately) most of the time nothing happens. The presented dataset on the other hand is designed to cover a wide range of different behaviors, including extreme cases and more subtle ones.

Still, future work will investigate how well the system performs during long-term operation in real-world setting, as in to the operational analysis of the audio component in [35], and at different sites (e.g. quiet station versus busy main station). Similar to [35], performance can be optimized by collecting site specific data and feedback by local surveillance experts. Another future direction is to extend the presented system to include higher level, semantic interpretation of the scene. Advancements in pose recovery in real-world conditions (e.g. [68]) could provide the necessary features, also allowing the detection of more subtle behavioral cues. Certainly, more sophisticated models will be needed to describe long term and complex relations between actions and events. The DBN could be adapted to model the time spent in each aggression state with a hidden semi-Markov model [69,70], similar to the train detectors in the current network that model the duration of the passing trains. Another approach is to use variable-length Markov models [71], which use arbitrary-order as opposed to the first-order Markov assumption, or Hierarchical HMMs where states at the higher levels generate sequences of lower level states, and a higher level state transition can only occur after a lower level sequence has finished.

9. Conclusions

This paper dealt with the detection of aggressive human behavior in complex, real-world scenarios. We used a DBN to estimate the latent variable, the aggression level, combining video, audio and contextual cues (interactivity, other objects). We also showed the benefit of combining the various cues, and the use of person-specific visual features derived from 3D person tracking.

Detection of aggressive behavior in complex real-world scenarios with multiple persons remains a challenging topic. Granted, we have not presented a system with the perfect prophecy capabilities of the mythological figure Cassandra. But with the performance achieved, we believe at least it will be more believable in signaling a modern-day equivalent of the Trojan horse.

Acknowledgments

We thank Wojtek Zajdel for his contribution to an early version of the CASSANDRA system. This research was supported by the Dutch Science Foundation NWO under grant 634.000.432 within the To-KeN2000 program. This research has also received funding from the European Community's Seventh Framework Programme under grant agreement number 218197, the ADABTS project.

References

[1] L. Dubbeld, The regulation of the observing gaze: privacy implications of video surveillance, (Ph.D. thesis), University of Twente, 2004.
[2] R. Baron, D. Richardson, Human Aggression, Springer, Berlin, 2004.
[3] S. Yudofsky, J. Sliver, W. Jackson, J. Endicott, D. Williams, The Overt aggression scale for objective rating of verbal and physical aggression, Am. J. Psychiatry 143 (2007) 35–39.
[4] D.M. Gavrila, Visual analysis of human movement: a survey, Comput. Vis. Image Understand. 81 (3) (2001) 231–268.
[5] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[6] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.
[7] A. Gupta, P. Srinivasan, J. Shi, L.S. Davis, Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 2012–2019.
[8] K.S. Huang, M.M. Trivedi, 3D shape context based gesture analysis integrated with tracking using OMNI video array, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, p. 80.
[9] N. Robertson, I. Reid, A general method for human activity recognition in video, Comput. Vis. Image Understand. 104 (2–3) (2006) 232–248.
[10] S.N. Vitaladevuni, V. Kellokumpu, L.S. Davis, Action recognition using ballistic dynamics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[11] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: real-time detection of violent crowd behavior, in: Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on, IEEE, 2012, pp. 1–6.
[12] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 3551–3558.
[13] A. Datta, M. Shah, N. Da Vitoria Lobo, Person-on-person violence detection in video data, Proc. Int. Conf. Pattern Recognit. 1 (2002) 433–438.
[14] A. Gupta, L.S. Davis, Objects in action: an approach for combining action understanding and object perception, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.
[15] Y. Luo, T.D. Wu, J.N. Hwang, Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks, Comput. Vis. Image Understand. 92 (2–3) (2003) 196–216.
[16] P. Peursum, G. West, S. Venkatesh, Combining image regions and human activity for indirect object recognition in indoor wide-angle views, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 1, 2005, pp. 82–89.
[17] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 1725–1732.
[19] B. Laxton, J. Lim, D. Kriegman, Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[20] B. Ni, S. Yan, A. Kassim, Recognizing human group activities with localized causalities, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1470–1477.
[21] Y. Zhou, S. Yan, T.S. Huang, Pair-activity classification by bi-trajectories analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[22] T. Yu, S. Lim, K. Patwardhan, N. Krahnstoever, Monitoring, recognizing and discovering social networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 1462–1469.
[23] D. Damen, D. Hogg, Recognizing linked events: searching the space of feasible explanations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 927–934.
[24] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.
[25] S. Calderara, R. Cucchiara, A. Prati, Bayesian-competitive consistent labeling for people surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 354–360.
[26] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2) (2008) 267–282.
[27] A. Mittal, L. Davis, M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene, Int. J. Comput. Vis. 51 (3) (2003) 189–293.
[28] J. Berclaz, F. Fleuret, P. Fua, Principled detection-by-classification from multiple views, in: Proceedings of the International Conference on Computer Vision Theory and Applications, vol. 2, 2008, pp. 375–382.
[29] R. Eshel, Y. Moses, Homography based multiple camera detection and tracking of people in a dense crowd, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[30] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, S. Maybank, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 663–671.
[31] S. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2006, pp. 133–146.
[32] K.N. Kutulakos, S.M. Seitz, A theory of shape by space carving, Int. J. Comput. Vis. 38 (3) (2000) 199–218.
[33] M. Liem, D.M. Gavrila, Multi-person tracking with overlapping cameras in complex, dynamic environments, in: Proceedings of the British Machine Vision Conference (BMVC), 2009, pp. 1–10.
[34] A.K. Nábělek, P.K. Robinson, Monaural and binaural speech perception in reverberation for listeners of various ages, J. Acoust. Soc. Am. 71 (5) (1982) 1242–1248.
[35] P.W. van Hengel, T.C. Andringa, Verbal aggression detection in complex social environments, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2007, pp. 15–20.
[36] K.R. Scherer, Vocal affect expression: a review and a model for future research, Psychol. Bull. 99 (2) (1986) 143–165.
[37] J.-C. Junqua, The Lombard reflex and its role on human listeners and automatic speech recognizers, J. Acoust. Soc. Am. 93 (1) (1993) 510–524.
[38] V. Cevher, A.C. Sankaranarayanan, J.H. McClellan, R. Chellappa, Target tracking using a joint acoustic video system, IEEE Trans. Multimedia 9 (4) (2007) 715–727.
[39] A. Noulas, B.J. Krose, On-line multi-modal speaker diarization, in: Proceedings of the Ninth International Conference on Multimodal Interfaces, 2007, pp. 350–357.
[40] I. Lefter, L.J. Rothkrantz, G.J. Burghouts, A comparative study on automatic audio–visual fusion for aggression detection using meta-information, Pattern Recognit. Lett. 34 (15) (2013) 1953–1963.
[41] T. Xiang, S. Gong, Incremental and adaptive abnormal behaviour detection, Comput. Vis. Image Understand. 111 (1) (2008) 59–73.
[42] Z. Yang, Multi-modal aggression detection in trains, (Ph.D. thesis), Technical University of Delft, 2009.
[43] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, N. Fakotakis, Fusion of acoustic and optical sensor data for automatic fight detection in urban environments, in: 13th Conference on Information Fusion (FUSION), IEEE, 2010, pp. 1–8.
[44] V.-T. Vu, F. Brémond, G. Davini, M. Thonnat, Q.-C. Pham, N. Allezard, P. Sayd, J.-L. Rouas, S. Ambellouis, A. Flancquart, Audio–video event recognition system for public transport security, in: IET Conference on Crime and Security, 2006, pp. 414–419.
[45] Y. Benezeth, P.M. Jodoin, V. Saligrama, C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurrences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2458–2465.
[46] O. Boiman, M. Irani, Detecting irregularities in images and in video, Int. J. Comput. Vis. 74 (1) (2007) 17–31.
[47] X. Wang, X. Ma, W. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2009) 539–555.
[48] M. Cristani, M. Bicego, V. Murino, Audio-visual event recognition in surveillance video sequences, IEEE Trans. Multimedia 9 (2) (2007) 257–267.
[49] W. Jiang, A.C. Loui, Audio-visual grouplet: temporal audio-visual interactions for general video concept classification, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2011, pp. 123–132.
[50] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, S. Theodoridis, Audio-visual fusion for detecting violent scenes in videos, Artif. Intell.: Theor. Models Appl., vol. 6040, Springer Berlin Heidelberg, 2010, pp. 91–100. Lecture Notes in Computer Science, isbn 978-3-642-12841-7.
[51] J. Lin, W. Wang, Weakly-supervised violence detection in movies with audio and video based co-training, Adv. Multimedia Inf. Process.-PCM 2009 (2009) 930–935.
[52] J. Nam, M. Alghoniemy, A.H. Tewfik, Audio-visual content-based violent scene characterization, in: Proceedings of the International Conference on Image Processing (ICIP), vol. 1, 1998, pp. 353–357.
[53] W. Zajdel, J.D. Krijnders, T. Andringa, D.M. Gavrila, Cassandra: audio–video sensor fusion for aggression detection, in: Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, 2007, pp. 200–205.
[54] Z. Zivkovic, F. van der Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, Pattern Recognit. Lett. 27 (7) (2006) 773–780.
[55] K. Kutulakos, S.M. Seitz, A theory of shape by space carving, Int. J. Comput. Vis. 38 (3) (2000) 199–218.
[56] J. Munkres, Algorithms for the assignment and transportation problems, J. Soc. Ind. Appl. Math. 5 (1) (1957) 32–38.
[57] J. Shi, C. Tomasi, Good features to track, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994, pp. 593–600.
[58] J.D. Krijnders, M.E. Niessen, T.C. Andringa, Sound event recognition through expectancy-based evaluation of signal-driven hypotheses, Pattern Recognit. Lett. 31 (2010) 1552–1559.
[59] T. Irino, R.D. Patterson, A time-domain, level-dependent auditory filter: the gammachirp, J. Acoust. Soc. Am. 101 (1) (1997) 412–419.
[60] B.C. Moore, B. Glasberg, A revision of Zwicker's loudness model, Acta Acustica United Acustica 82 (2) (1996) 335–345.
[61] R. Hut, M.M. Boone, A. Gisolf, Cochlear modeling as time–frequency analysis tool, Acta Acustica United Acustica 92 (4) (2006) 629–636.
[62] P. van Hengel, J.D. Krijnders, A comparison of spectro-temporal representations of audio signals, IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2) (2014) 303–313.
[63] R. McAulay, T. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, Proc. Int. Conf. Acoustics Speech Signal Process. 34 (4) (1986) 744–754.
[64] A.S. Bregman, Auditory Scene Analysis, The MIT Press, 1990.
[65] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explorations 11 (1) (2009) 10–18.
[66] Z. Ghahramani, An introduction to hidden Markov models and Bayesian networks, IJPRAI 15 (1) (2001) 9–42.
[67] X. Boyen, D. Koller, Tractable inference for complex stochastic processes, in: Proceedings of the UAI, 1998, pp. 33–42.
[68] M. Hofmann, D.M. Gavrila, Multi-view 3D human pose estimation combining single-frame recovery, temporal integration and model adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2214–2221.
[69] K.P. Murphy, Hidden semi-markov models (hsmms), Technical Report, MIT AI Lab, 2002.
[70] Q. Shi, A. Canberra, L. Wang, C. Nanjing, L. Cheng, A. Smola, Discriminative human action segmentation and recognition using semi-Markov model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[71] A. Galata, N. Johnson, D. Hogg, Learning variable length Markov models of behaviour, Comput. Vis. Image Understand. 81 (3) (2001) 398–413.