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/

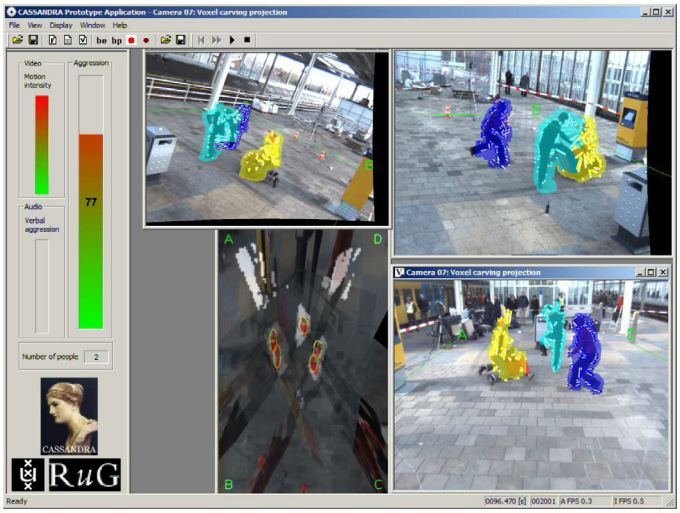

Fig. 1. A screenshot of the CASSANDRA prototype application. In the application three windows show camera images in which detected persons are annotated by a (random) color. A fourth window in the bottom left displays the top-down projection (constructed from image homography) with voxel carving results 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