Anti-social Behavior Detection in Audio-Visual Surveillance Systems

Jogile Kuklyte[1],Philip Kelly[1],Ciarán Ó Conaire[1],Noel E. O'Connor[1] and Li-Qun Xu[2]

[1] CLARITY: Center for Sensor Web Technologies, Dublin City University, Ireland
[2] Visual Computing and Multimedia Understanding, British Telecom, UK

Abstract. In this paper we propose a general purpose framework for detection of unusual events. The proposed system is based on the unsupervised method for unusual scene detection in web–cam images that was introduced in [1]. We extend their algorithm to accommodate data from different modalities and introduce the concept of *time-space blocks*. In addition, we evaluate early and late fusion techniques for our audio-visual data features. The experimental results on 192 hours of data show that data fusion of audio and video outperforms using a single modality.

1 Introduction

Starting from two cameras in Trafalgar Square, London in 1960 and a much bigger construction called "Ring of Steel" consisting of thousands of cameras placed around London in 1990's, video surveillance systems were introduced to assist the police and security guards in preventing crime.

The number of cameras is growing rapidly. In 2002 the number of cameras in the UK was around 4.2 million, which is approximately one camera for every fourteen people[3]. Difficulties can occur if one operator has to monitor multiple camera views at the same time, as the optimal concentration span for a person is about 25-30 minutes[4]. Taking into account all these facts, a new generation of surveillance systems with real–time data processing is needed where most of the work load would be done by computer.

There are already commercially available surveillance systems with video content analysis adapted for different tasks, but such systems are quite expensive, often need a professional to set it up and they are optimized for specific tasks in specific environments. Research effort is underway to make systems cheaper by integrating different sensors as well as improving how they function. Sensors such as audio, motion, multi spectral (thermal and infrared) cameras could increase confidence in the results and add a different perspective on the events happening in the scene.

Our proposed approach employs acoustic and visual data to detect unusual situations, for example fighting in a usually quiet corridor. More than a week

[3] from the M. McCahill and C. Norris report, June 2002; http://www.urbaneye.net

[4] from "People in control: human factors in control room" by J. M. Noyes and M. Bransby, 2001, pages 40-41

of audio-visual data (24 hours a day) was collected from an indoor environment (example data can be seen in figures $3c$ to $3f$). Overlapping time-space block features were calculated for time stamped audio and video data. Unsupervised classification with a Euclidean similarity measure was applied and two different ways of fusing the two modalities – early and late fusion – were tested.

This paper is organized as follows: Section 2 outlines previous work in this area and gives an overview of the proposed system. Section 3 details the proposed abnormal behavior detection system and the components that constitute the key contributions in this paper. Section 4 provides experimental results of the system framework from a specific application scenario. Conclusions and directions for future work are described in section 5.

2 Prior Work

With the large number of surveillance cameras now in operation, both in public spaces and in commercial centers, significant research efforts have been invested in attempts to automate surveillance video analysis [2]. Breitenstein et al proposed the algorithm for novelty detection in video data [1]. Using simple features and a clustering algorithm they detect unusual scenes (abnormal activity). In their method everything that was seen during training period is defined as "usual". This approach works well with low frequency web-cam images.

Some work has been done in analyzing audio alone. Clave et al [3] described a supervised audio event detection model which detects gunshots. Atrey et al [4] experimented with modeling events using Gaussian mixture models with four different audio features - Zero Crossing Rate (ZCR), Linear Prediction Coefficients (LPC), Linear Frequency Coefficients (LFC), Linear Frequency Cepstrum Coefficients (LFCC). They used a single microphone to detect a set of events such as shouting, knocking, talking, and footsteps (walking and running).

Both [5] and [6] argue that the integration of video technology with sensors and other media streams will constitute the fundamental infrastructure for new generations of multimedia surveillance systems. The goal is to boost detection results from any one modality by combining analysis results from multiple complementary modalities. A framework for transport security was proposed in [7]. They combined face detection and tracking with audio event detection to perform audio-video scenario recognition and were able to successfully recognize several scenarios. A multimodal approach for detecting events in meeting environments was proposed in [8]. They showed that adding video and localization information to acoustic information improved the detection of some events. Another audio visual event recognition system was proposed in [9]. They implemented The Time Adaptive Mixture of Gaussians (TAPPMOG) probabilistic method. Smeaton & McHugh [10] used simple features to detect audio activity in a computer room. They examined if audio analysis could be employed to assist visual event detection system and whether simple low level features produce reliable results. They showed that audio data can be a significant aid in surveillance and security applications.

3 Abnormal Activity Detection

In this work, we present a significant extension to the work first proposed by Breitenstein et al [1], which describes an unsupervised data-driven technique for the detection of unusual events from a sequence of static web-cam images. Using a frame rate as low as 3 images per minute they were able to detect unusual scenes such as crowds of people walking on the road during a festival, or roadworks in a city environment.

This paper extends the work with two important contributions. The first involves generalizing the algorithm from being applicable to discrete images or frames taken every few seconds, to continuous data streams from a variety of sensor types. This process is achieved via the use of *time–space block* feature vectors, as described in section 3.1. The second contribution of this paper involves the integration of data fusion methodologies into the abnormal event detection framework outlined by [1]. Two data fusion methodologies are described in section 3.2 and subsequently compared and evaluated in section 4.

A limitation of Breitenstein et al's proposal lies in the fact that only discrete webcam images were used as inputs to their algorithm. The first extension of this algorithm allows the classification of arbitrary data streams with higher frame rates in real-time via the use of simpler features. We also assume that events span a duration of time and implement the *time–space block* feature vectors. The *time–space block* size depends on the events to be detected, for example antisocial events tend not to happen in a single frame. The second contribution in this work extends the algorithm to make it applicable to more than one input datastream. In our approach information from both video and audio data features are employed for unusual event detection. The aim is to enrich the definition of an event from being solely image feature based to becoming an audio-visual event – thereby increasing the robustness and scope of our detections.

3.1 Time-Space Blocks

In the proposed approach, for every examination of the data a feature vector of length a is created: $T_f = \{f_i, f_{i+1}, \cdots, f_{i+a}\}$. Each element f_i in the feature vector T_f represents m seconds of activity in the data stream. This element can be obtained by either sampling the stream at a single point in time, by averaging the data within that m time frame slice or by applying more sophisticated analysis. The contents of the data buffer is analyzed every $b \times m$ seconds (see figure 1), meaning that in the worst case scenario the detection of events is only $b \times m$ seconds late. For both audio and video streams a multitude of different feature types of varying properties and complexities could have been chosen to represent the underlying data within T_f ([10],[11],[12],[13],[1]). However, in order to process the data in real–time we choose to use low–complexity features.

For *Audio Feature Vectors* we used the Root Mean Square (RMS) measure of the audio mean power within the whole of the timeslice m (figure 1). We employ the use of RMS in this work as it provides a good representation–versus–complexity tradeoff. In essence, RMS can be viewed as a statistical measure of

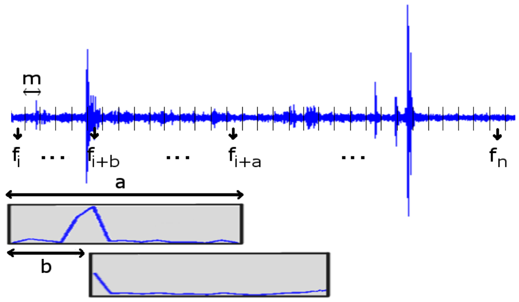


Fig. 1. Time-space blocks - audio example. At every b interval, the time segments m are transformed into features f_i that are grouped into vectors of size a.

the magnitude of a varying audio waveform and distribution of this measure over time quite accurately can distinguish between different events. For *Video Feature Vectors*, the feature chosen was the number of foreground pixels per frame that occurs nearest to the time m. Foreground pixels were calculated using a frame difference method $|fr_{m+1} - fr_m| > Th$, where fr_m is a frame at the time m and Th represents the threshold for the foreground. During *Synchronisation* we assume that the data streams are temporally aligned. In our experiments, synchronization between two data streams, one video and one audio, is achieved via the use of a single audio-visual camera with MPEG-4 transmission capabilities.

3.2 Data Fusion

We want to improve reliability and to detect events that single modality data analysis would not be able to achieve. Such an example could be verbal abuse, that would not be detected with video analysis, or events such as stealing a neglected item that would not be possible to detect solely using audio data. Although we have used video and audio modalities in this work, it should be noted that this framework is flexible and that the main idea of integrating fusion in to our event detector model is to lay the groundwork for the fusion of many different numbers of heterogeneous or homogeneous modalities.

Audio–visual fusion can be accomplished in different levels. We explore two of them – feature level and decision level fusion. Feature level (early) fusion is carried out by combining audio and video features to construct joint feature vectors. These feature vectors are then used to classify events in the same manner as if they were from a single input modality. When decision level (late) fusion is applied the results from each modality are combined afterwards.

One of the main issues involved in early fusion is how to scale and weight the features coming from different sensor modalities. In our experiment we normalized all the feature values v_i^x to unity:

$$\hat{V}_x = \{\hat{v}_0^x, \hat{v}_1^x, \ldots, \hat{v}_n^x\}, where\ \hat{v}_i^x = \frac{v_i^x}{\hat{v}_{max}^x} \qquad (1)$$

In addition different modalities can have different amounts of information that they are carrying and to balance between them we need to choose weights. Weighting of the features from different modalities was performed as follows:

$$V_{final} = cat\{\alpha\hat{V}_x, \beta\hat{V}_y, \ldots, \gamma\hat{V}_z\} \tag{2}$$

where *cat* is the concatenation of the vectors and $\|\alpha + \beta + \ldots + \gamma\| = 1$. In this work we implemented both early and late fusion to be able to compare the results. Results are presented in section 4. For early fusion we implemented the techniques as described above. For late fusion, we analyze two different methodologies: performing a binary AND and a binary OR operators on anomalous events detected by the audio and video detectors separately.

3.3 Classification

During the training stage an agglomerative clustering algorithm [1] is applied. In this technique, q model clusters are used to represent usual scenes (in our experiments $q = 100$ due to processing time constraints). During the training stage, every observation is saved until the maximum number of clusters is reached. After this stage, outliers from the models (or unusual events) are found using dynamic threshold which is calculated via meaningful nearest neighbor technique (where the percentage of the probability distribution, p_{alarm}, used in [1] was set to 0.01 or 1%). The q model clusters are constantly updated using a cluster weighting technique (see [1] for more details).

4 Experimental Results

We collected data with an AXIS 212 audio-visual camera[5] placed in the corridor outside a research laboratory, figure 2. The environment chosen is challenging because of the big window along the corridor which introduces shadows and varying lighting conditions, weather surroundings, different time of day and ambient noise from outside the window picked up by audio sensor. The usual events in the scene consists of a steady stream of people walking through the scene, talking on mobile phone by the window, meeting in groups etc.20 potential antisocial events were performed by actors (see table 2) and labeled for evaluation. The audio and video features were calculated every second (m=1s) in order to preserve sufficient details for the events whilst still being able to process the data in real time. Other parameters are set as $a = 15$, $b = 5$, $\alpha = 0.5$ and $\beta = 0.5$, where a is a dimension of feature vector and b is a shift size, α and β are the audio and video weights for the early fusion that were chosen empirically. We evaluated results by the number and type of events that were detected.

As can be seen from the results in table 1 total number of detected events is significantly greater than the acted events. The ground truth antisocial scenarios all together were 24.4 minutes long (see table 2) which is around 0.2% of the

[5] http://www.axis.com/files/datasheet/ds_212ptz-v_34051_en_0812_lo.pdf

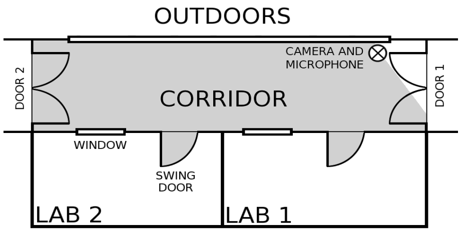
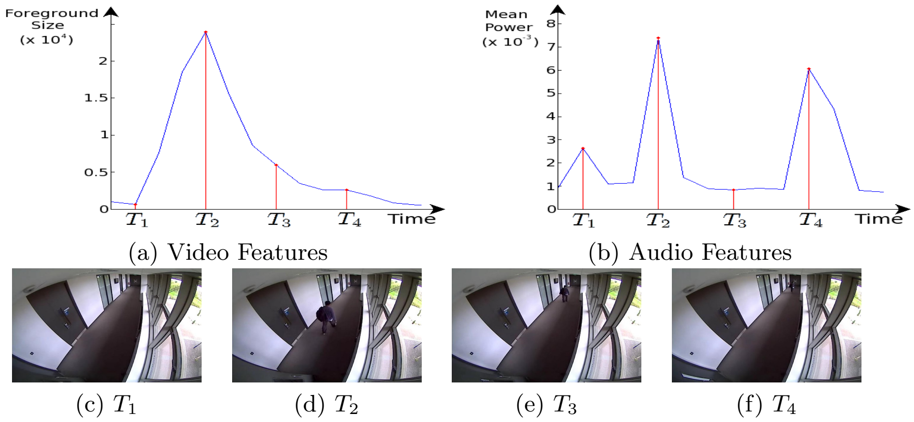


Fig. 2. The capture environment. The camera location is indicated by $\otimes$.



(a) Video Features

(b) Audio Features

(c) T_1

(d) T_2

(e) T_3

(f) T_4

Fig. 3. Example 15 seconds of audio and video feature frame of a person walking through the corridor. It starts with opening of the doors (T_1). As a person walks away from the camera (T_2 to T_3) the foreground size becomes smaller. The audio feature graph shows the peaks where the sound is produced by opening and closing of the door nearest to the camera (bottom left in the images T_2 and T_4).

experimental data. The most of the events that were falsely detected as unusual were opening and closing the door close to the camera that were deemed as significantly different to the 100 background models of the scene. Also there were some random lightning changes that were falsely detected as events, as well as increase in audio noise when the grass mower was cutting the grass. After reviewing the results, all these events could be manually labeled as usual and incorporated into the usual event database. Most of the acted events (95%) were detected using late fusion of audio and video modalities by operand OR. It can also be seen that number of events detected by this method is also the highest. Early fusion of audio and video modalities detected 90% of acted events with half of the amount of total events detected by previously mentioned technique. As we can see from the table the overlap between events detected by audio and video is about 75%, so there are events that were detected by a single modality.

The only event that was not detected by any modality was a person standing

Sensor	modalities	N_t	N_p
AU	*audio*	285	17
VI	*video*	304	16
FU	*AV early fusion*	280	19
AND	*AV late fusion AND*	120	15
OR	*AV late fusion OR*	551	18

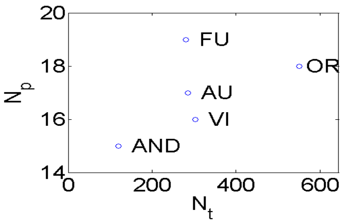


Table 1. Cumulative system performance: where N_t = total number of events detected and N_p = number of events detected from the list of performed events

Scenario	duration	AU	VI	FU	AND	OR
Two people fighting	15sec	✓	✓	✓	✓	✓
Phone ringing	13sec	✓	✗	✓	✗	✓
Someone putting a poster on the wall	4min 15sec	✓	✓	✓	✓	✓
Shouting and fighting	15sec	✓	✓	✓	✓	✓
Climbing on the window sill and jumping off	10sec	✓	✓	✓	✓	✓
Running to the lab	30sec	✓	✗	✓	✗	✓
Tearing the poster from the wall	40sec	✓	✓	✓	✓	✓
Bullying/intimidation	1min 05sec	✓	✓	✓	✓	✓
Waving to the camera	15sec	✓	✓	✓	✓	✓
Screaming	2sec	✓	✓	✓	✓	✓
Attempt to enter a laboratory	30sec	✓	✓	✓	✓	✓
Football in the corridor	1min 05sec	✓	✓	✓	✓	✓
Bringing in a ladder, climbing on it	5min 30sec	✓	✓	✓	✓	✓
Shouting	2sec	✓	✓	✓	✓	✓
Arguing near the lab door	10min 50sec	✓	✓	✓	✓	✓
Running through the corridor	10sec	✓	✓	✓	✓	✓
Someone standing on their head	30sec	✗	✗	✗	✗	✗
Leaving something in the corridor	9sec	✗	✓	✓	✗	✓
Removing something from the corridor	30sec	✓	✓	✓	✓	✓
Activity outside the window	30sec	✗	✗	✓	✗	✗

Table 2. List of scenarios and detection results

on his head. Taking into account the features that we used, the event itself is hard to separate from a person standing on his feet, which is a usual event. Events that were detected by all combinations of sensors were the events that could be classified as antisocial. These events includes fighting, tearing the poster from the wall, playing football in the corridor, arguing etc. Leaving an unattended object in the corridor was not detected by audio, while a phone ringing and running to the lab was not detected by video but detected by the other modality. A trolley being dragged outside the window was detected only by early fusion.

5 Conclusion and Future work

In this work we develop a scalable framework, that is scalable to a variety of heterogeneous or homogenous data inputs, and employ it for the detection of unusual events in audio-visual data streams. Although video is very important for surveillance, we highlight importance of audio and show that the fusion of the two features outperforms any single modality. We believe that the results can be improved by introducing audio frequency features and more complex video features. Moreover adding confidence measures and more sophisticated weighing methods should help to improve the fusion results.

Acknowledgements

This work is supported by BT Group plc and the Irish Research Council for Science, Engineering and Technology: Funded by the National Development Plan and by Science Foundation Ireland under grant 07/CE/I1147.

References

1. M. D. Breitenstein, H. Grabner, and L. V. Gool. Hunting nessie – real-time abnormality detection from webcams. In *VS*, Kyoto, Japan, October 2009.
2. T. Ko. A survey on behavior analysis in video surveillance for homeland security applications. In *AIPR*, pages 1–8, 2008.
3. C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *ICME*, pages 1306–1309, Amsterdam, Netherlands, July 2005.
4. P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio based event detection for multimedia surveillance. In *ICASSP*, pages 813–816, Toulouse, May 2006.
5. R. Cucchiara. Multimedia surveillance systems. In *VSSN*, pages 3–10, 2005.
6. W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3):334–350, 2004.
7. V. T. Vu, F. Bremond, G. Davini, M. Thonnat, Q. C. Pham, N. Allezard, P. Sayd, J. L. Rouas, S. Ambellouis, and A. Flancquart. Audio-video event recognition system for public transport security. In *ICDP*, pages 470–475, London, June 2006.
8. C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J. R. Casas. Audiovisual event detection towards scene understanding. In *CVPR*, pages 81–88, 2009.
9. M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transactions on Multimedia*, 9(2):257–267, 2007.
10. A. F. Smeaton and M. McHugh. Towards event detection in an audiobased sensor network. In *VSSN*, pages 87–94, Singapore, November 2005.
11. C.Couvreur P. Gaunard, C. G. Mubikangiey and V. Fontaine. Automatic classification of environmental noise events by hidden markov models. *IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pages 3609–3612, 1998.
12. L. Khan G. Lavee and B. Thuraisingham. A framework for a video analysis tool for a suspicious event detection. In *MDM*, Aug. 2005.
13. C. Bailey H. Meng, N. Pears. A human action recognition system for embedded computer vision application. In *ECV*, Minneapolis, USA, 2007.