# PhD Forum: Investigating the Performance of a Multi-modal Approach to Unusual Event Detection

Jogile Kuklyte, Philip Kelly and Noel E. O'Connor
CLARITY: Center for Sensor Web Technologies, Dublin City University, Ireland
Email: jogilek@eeng.dcu.ie

*Abstract*—**In this paper, we investigate the parameters underpinning our previously presented system for detecting unusual events in surveillance applications [1]. The system identifies anomalous events using an unsupervised data-driven approach. During a training period, typical activities within a surveilled environment are modeled using multi-modal sensor readings. Significant deviations from the established model of regular activity can then be flagged as anomalous at run-time. Using this approach, the system can be deployed and automatically adapt for use in any environment without any manual adjustment. Experiments carried out on two days of audio-visual data were performed and evaluated using a manually annotated ground-truth. We investigate sensor fusion and quantitatively evaluate the performance gains over single modality models. We also investigate different formulations of our cluster-based model of usual scenes as well as the impact of dynamic thresholding on identifying anomalous events. Experimental results are promising, even when modeling is performed using very simple audio and visual features.**

## I. Introduction and Related Work

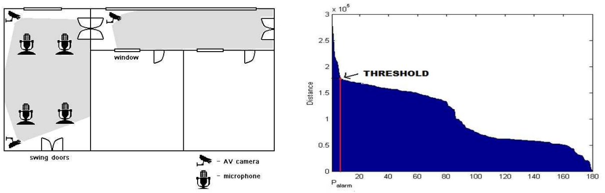
In this work, we describe and further investigate our previously presented system (see [1]) that can relieve the workload on human surveillance operators, by automatically identifying anomalous or suspicious events happening in the surveilled areas. A key feature of our approach is that it uses audio as well as visual data. Audio sensors are not commonly used in traditional surveillance systems due to the overlap phenomena [2] but we propose that audio data can provide useful information about the environment under surveillance. e.g. the level of noise, the location of the sound source, and decoded sound frequencies can be used to recognize the source of the sound [3]. Furthermore, according to Pavlidis et al [4], price is one of the foremost issues preventing the adoption of new technologies in surveillance systems, yet audio sensors are relatively cheap and require little transmission bandwidth. In typical analytic systems, events are manually predefined in advance [5]. We leave event recognition for the next stage of our work and first focus on detecting unusual scenes. Other unsupervised learning approaches such as [6] learn usual movements during the period of time and use selected data to train the system. This reduces manual labeling of training data, but has no capability to adapt to changing environment and the training data has to be carefully selected. For robust surveillance, the learning should not stop after the training phase due to the typically dynamic nature of any environment e.g. cyclic light changes, weather and seasonal changes.

## II. Proposed Surveillance System

In our previous work [1] we described a multi-modal capture environment spanning two corridors, illustrated in figure 1(a). The analysis system is composed of two main stages: *training* and *classification*. Training is a continuous task that is undertaken prior to classification for a period of time, but also continues to update the model during the classification phase. During training, typical scenes of a surveilled area are modeled using agglomerative clustering algorithm using low level visual and audio features. For video features, foreground size of each frame is calculated from the foreground mask which is acquired using a Mixture of Gaussians (MoG) method proposed by KaewTraKulPong *et al* [7]. For acoustic features audio energy level is used. Its variation over time provides a simple but descriptive signature of the scene. In both modalities the temporal feature vectors are created using a sliding rectangular window applied to the sequential data. All the feature vectors are normalized to the [0..1] interval by applying equation $x_{norm} = \frac{x-min}{max-min}$, where $max$ and $min$ values are calculated from the dataset for each modality separately. Audio-visual fusion is obtained by concatenating normalized feature vectors and giving them equal weights. The algorithm requires a number of clusters $N$ to be set in advance. We investigate the affect of changing the number of clusters in the experiments section. During classification, a Nearest Neighbor (NN) algorithm is applied to determine whether the new instance represents an unusual event or whether it belongs to the background activities. We apply an adaptive thresholding method proposed by Breitenstein *et al* [8]. The method accumulates shortest distances between the trained model and data over time. The threshold is then chosen from the sorted collection of these shortest distances with the help of parameter $P_{alarm}$ – see figure 1(b). The figure shows a collection of the shortest distances $D$ collected over a period of time ($t$ number of frames). Here, shortest distances are sorted in a descending order and a parameter $P_{alarm}$ defines the position in the collection $D$ from where the threshold will be acquired based on the formula $THRESHOLD = D(t * P_{alarm})$.

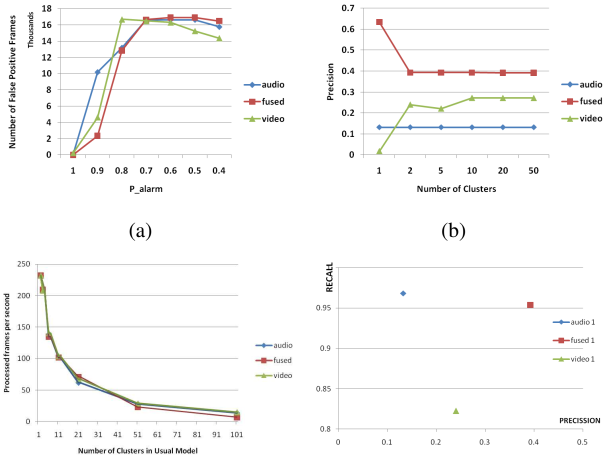
## III. Experiments and Results

In our experiments here we focus on the selection of parameters that would bring the most benefit with the least computational complexity. Experiments are carried out on surveillance data collected in the indoor environment. We used one day of data for training and we assume that it consists



(a) (b)

Fig. 1. (a) Capture set-up showing the location of multiple cameras and microphones across two rooms in an indoor location; (b) Illustration of adaptive thresholding.



(a) (b)

(c) (d)

Fig. 2. (a) Number of false positives vs. choice of adaptive thresholding parameter; (b) Overall precision vs. number of clusters used to represent *usual* events; (c) Processing time (frames/sec) vs. number of clusters used; (d) Overall precision & recall for audio, video and fusion.

mostly of usual scenes. Six unusual events were deliberately performed during the second day used for testing. The list of activities is composed of (i) acoustic events, i.e. playing guitar, screaming and shouting, (ii) visual events i.e. stealing a fire extinguisher and (iii) audio-visual events i.e. banging on the doors, playing football & fighting.

### A. *Adaptive thesholding*

By running experiments with different $P_{alarm}$ values we found that the best results were given when the parameter was set to $0.9$ – see figure 2(a). Because we chose the threshold value from the list of shortest distances, most of the distances relate to the usual model, thus in theory we should be able just to choose the highest value from the shortest distances and use it as a threshold. But as our training data was unsupervised, and there is a possibility that some outliers pollute this data, we have to discard some of the top shortest distances from the list.

### B. *Clusters and processing time*

From figure 2(b), it can be seen that there is no improvement in detection results after using more than two clusters to represent unusual events, except in terms of processing time, see figure 2(b). This is due to the fact that we deliberately use simple descriptors.

### C. *Fusing audio and video*

In our previous preliminary work we confirmed the hypothesis that early fusion of visual and acoustic information gives better results compared with late fusion techniques and single modalities [1]. Figure 2(d) provide further evidence for this. Fusion improves the precision over using video only by 15% and over using audio only by 25%, while recall for video only is improved by 13% and is only 2% smaller for fusion than for using audio only.

## IV. Conclusion and Future Work

The effect of varying the fixed parameters for a multi-modal approach to detecting anomalous events proposed in [1] was investigated. The number of clusters and threshold used were investigated in this paper. Low complexity descriptors were deliberately used in order to investigate the potential for the algorithm to run in real-time or to ease implementation directly in camera hardware. In the next stage of this work we plan to investigate different descriptors for anomalous event detection to investigate whether the advantages of using more complex features would outweigh the disadvantages of increased processing power demands. We will also investigate classification of the detected events into categories. At this stage, more sophisticated features will need to be implemented to distinguish between more categories of events, but the processing will be carried out only on the detected events and may be thus affordable.

## Acknowledgment

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] J. Kuklyte, P. Kelly, C. Ó Conaire, N. E. O'Connor, and L. q. Xu, "Anti-social Behavior Detection in Audio-Visual Surveillance Systems," in *PRAIxHBA*, no. December, 2009.

[2] T. Butko, C. Segura, C. Nadeu, J. Hernando, and J. R. Casas, "Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion," *Accepted to Interspeech*, pp. 2–5, 2009.

[3] D. Orla, S. Marlow, M. Noel, O. Noel, and S. Alan, "Road Traffic Monitoring using a Two-Microphone Array," in *Audio Engineering Society Convention 118*, 2005.

[4] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, "Urban Surveillance Systems: From the Laboratory to the Commercial World," *Proceedings of IEEE*, vol. 89, no. 10, pp. 1478–1497, 2001.

[5] M. Dikmen, H. Ning, D. Lin, L. Cao, and V. Le, "Surveillance event detection," *In TRECVID Video Evaluation Workshop*, 2008.

[6] T. Nanri and N. Otsu, "Unsupervised abnormality detection in video surveillance," in *IAPR Conference on Machine Vision Applications*, no. 574-577. Citeseer, 2005, pp. 574–577.

[7] P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real- time Tracking with Shadow Detection," *2nd eouropean workshop on advanced video-based surveillance systems*, pp. 1–5, 2001.

[8] M. D. Breitenstein, H. Grabner, and L. V. Gool, "Hunting nessie - real-time abnormality detection from webcams," *Workshop On Visual Surveillance*, Oct. 2009.