Aggression Detection in Urban Environment based on Audio Analysis

Review Report

Author: Edward Fleri Soler
edward.fleri.14@um.edu.mt

Supervisor: Dr George Azzopardi
george.azzopardi@um.edu.mt

ABSTRACT

In this project computer vision techniques are introduced into the field of security and surveillance, in an attempt to design and implement an automatic aggression detection system through the analysis of audio from urban environments. The proposed system differs in approach to similar applications in literature, through the incorporation of visual analysis techniques as opposed to the direct analysis of low-level audio features. This novel, strategic approach is motivated by the demand for a higher robustness to noise, as well as the improved ability to deal with both impulsive and sustained events of interest. The MIVIA Audio Events dataset was selected for the testing and thorough evaluation of the system, with the functionality of the proposed pipeline being confirmed, illustrating its ability to confidently detect and classify events of importance within highly suboptimal conditions.

Keywords

Event detection, Audio analysis, Visual analysis, Localisation, Classification, Signal-to-Noise ratio

1. INTRODUCTION AND BACKGROUND

With advancements in technology, reductions in production cost and a greater demand than ever, surveillance systems are being installed in parks, squares, stations and public spaces around the world. These systems provide peace of mind to communities, aid law enforcement in the prevention of crime and promote public safety [1]. Visual surveillance, in the form of closed-circuit television (CCTV) cameras, is by far the leading means of security surveillance. However, this approach is by no means flawless. Visual surveillance methods may fail to identify threats in crowded environments due to the partial or total obscurity of areas of interest. Current technologies also require manual detection of threats, and are thus restricted by the man power available to analyse footage and other sensory information.

Audio sensors are generally cheaper to produce and require less processing power than CCTV cameras, while overcoming illumination issues, deeming them equally suitable during day and night time [2]. Furthermore, the omnidirectionality of modern microphones trumps the limited field of view associated with visual sensors. This not only allows for the detection of events over a larger area, but also facilitates the identification of important events, such as gunshots and screams, which hold no visual signature [3]. Within hectic environments, automated audio analysis is a far more effective approach, overcoming the above mentioned shortcomings of its visual counterpart.

The greater majority of audio analysis systems follow a pipeline based on the direct extraction of low-level audio features and characteristics, for the identification and classification of events of interest. In [4], Clavel et al. make the first step towards their vision of a multimedia abnormal event detection system, by focusing on the detection of gunshots. This procedure begins with the segmentation of audio into partially overlapping frames of 20ms in temporal width, before exploiting characteristic energy, spectral and Mel-Frequency Cepstral Coefficient (MFCC) audio features, for the generation of feature vectors. These features are then abstracted to a higher level, through the application of appropriate models to generate event class descriptors, capable of summarising the characteristic features of a class of audio events. Classifiers are then trained upon these models, in order to sequentially analyse and classify test input frames.

Carletti et al.'s aggression detection system [5] employs an approach analogous to that of Clavel et al. [4], with the addition of the Bag-of-Words model for the high-level abstraction and description of events. The simplicity of, and the noticeable improvements brought about by, the Bag-of-Words model were similary realised by Foggia et al. during the implementation of their road surveillance system [2], and aggression detection system [6], with functionalities akin to those of the proposed pipeline.

Dennis et al. [7] provide inspiration for the adaptation of computer vision techniques, as opposed to the typical approaches inspired by speech recognition, for the detection and classification of audio events. The authors employ the descriptive power of spectrogram-like representations for the depiction of the rich visual signatures of both impulsive and sustained events. The capability of this approach was reconfirmed by Dennis et al. [8], as well as Foggia et al. [9], whereby the highly satisfactory results showcased its capacity at dealing with suboptimal conditions as well as events of different temporal diameter.

In this project we shall explore the idea of audio surveillance in urban environments, focusing on the identification of audio signals relating to the breaking of glass, gunshots and screams in an off-line setting. The Mivia Audio Events Dataset, compiled by Foggia et al. for the testing of their approach to aggression detection [6], shall be employed. Consisting of over 28 hours of audio tapes, this dataset holds 18,000 events of aggression superimposed with vary levels of background noise typical of urban environments.

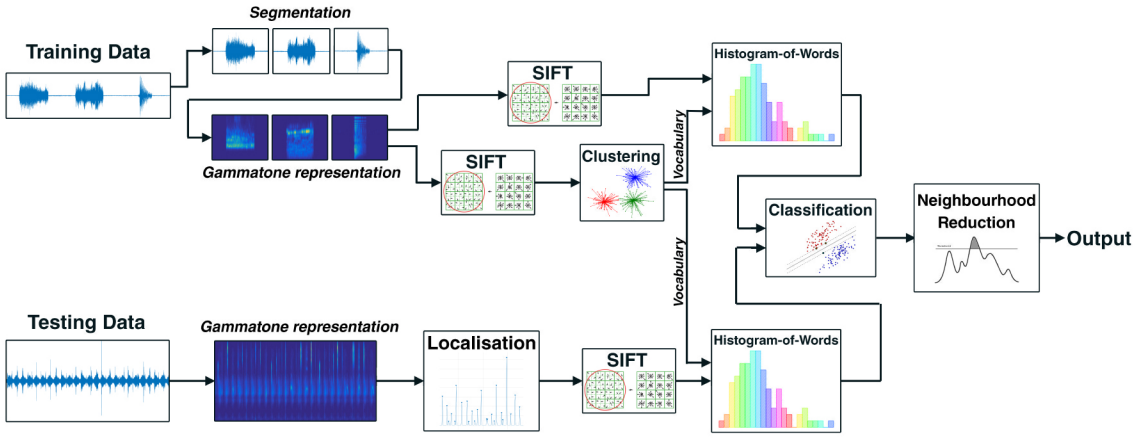

Figure 1: Pipeline of the proposed system

2. AIMS AND OBJECTIVES

The aim of this project is to devise a system which automatically detects aggressive events in urban environments through the analysis of audio signals. This goal shall be reached through the following intermediate objectives:

1. Conversion of audio files to visual representations
2. Image features extraction and vocabulary generation
3. Training of classifiers on event image features
4. Localisation and classification of test data events
5. Comprehensive system evaluation

3. DESIGN

3.1 Data Handling and Audio Representation

Figure 1 above outlines the main procedures involved in the proposed pipeline. The first procedure involves the segmentation of audio input so as to generate a labelled set of event and background noise snippets. Next, each snippet is transformed into its respective time-frequency representation, as may be observed in figure 2. The employment of different representations yields different results, deeming the image representation parameter to be an important one.

3.2 Feature Extraction and Description

The SIFT algorithm [10], routinely used within literature, is implemented for the extraction of image features, known as keypoints. A set of keypoints are extracted from each snippet, including background noise, before undergoing clustering through the K-Means algorithm to form a dictionary of size K. The resulting clusters form the code book on which all image descriptors shall be based from this point on.

The incorporation of spatial data into image descriptors is of key importance, as each class of events holds a characteristic energy signature at specific temporal and spectral regions. Spatial pyramids are therefore employed so as to relate each keypoint to a spatial zone within the image. The SIFT algorithm is once again implemented to extract keypoints from each of the spatial zones, before abstracting these features into histogram-of-words, as dictated by the Bag-of-Words model.

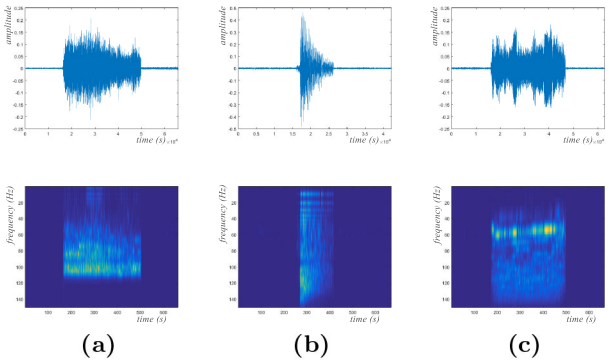

Figure 2: Waveform and corresponding gammatonegrams of (a) breaking glass, (b) gunshot & (c) scream

3.3 Localisation

As opposed to the typical windowing method, the proposed pipeline follows a hierarchical detection approach, locating suspected events prior to classification. Localisation is performed by applying a line detection algorithm, tuned to respond to vertical lines, to the visual representation of an input test audio file. This algorithm returns an energy response in regions holding vertical lines, while ignoring non-linear textures or lines at other orientations. The response is then summed across the frequency domain, producing a plot of total response at each temporal unit. Peaks within this plot are taken to be the center of suspected events, while the troughs on either side are respectively taken to be the start and end times. The test audio file is then segmented according to the located start and end times, with a histogram-of-words being computed for each suspected event segment.

3.4 Classification

Training and testing data are now represented in the same manner, allowing for Support Vector Machine (SVM) classification to take place. Each suspected event is assigned a label relating it to one of the event classes, or alternatively dismissing it as noise. A process known as neighbourhood reduction is then applied to analyse the localisation, or classification, scores of each of the suspected events, marking any suspected events below a threshold as noise.

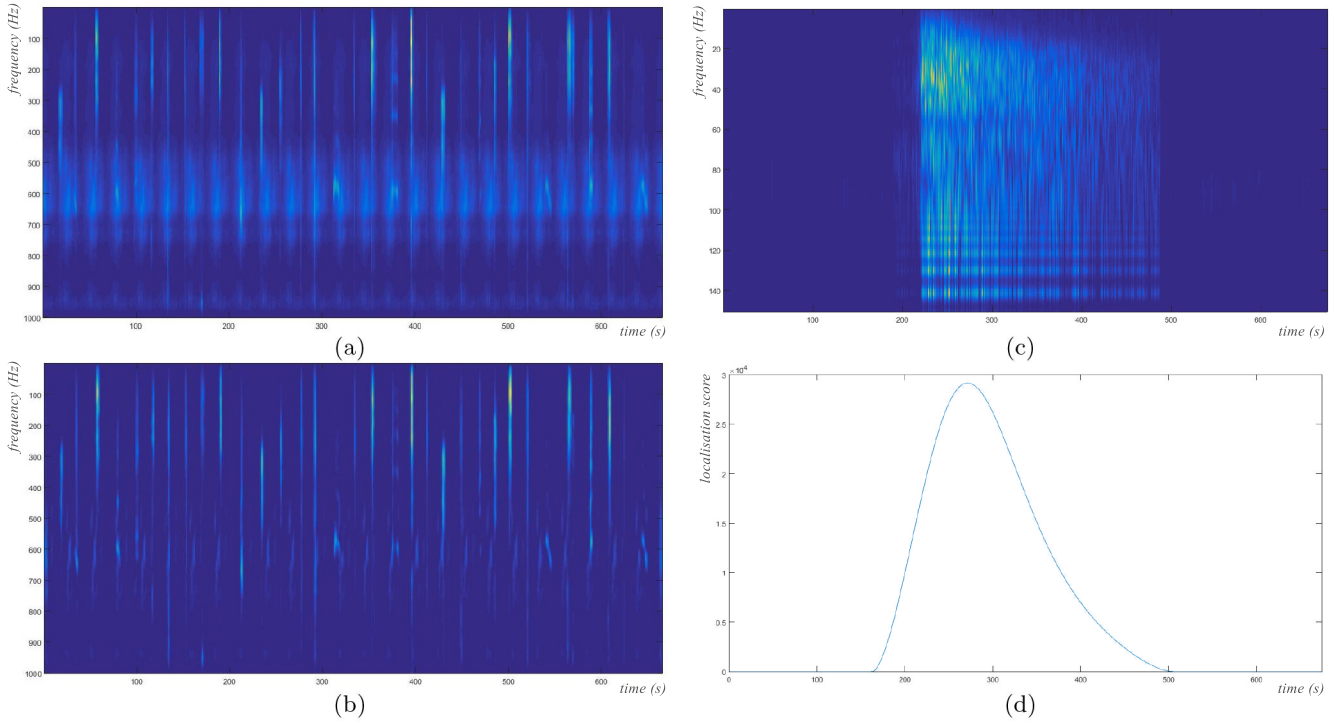

Figure 3: (a) Gammatonegram of audio file consisting of 33 events at a 10dB SNR level. (b) Response of the vertical line detection algorithm applied to the audio file gammatonegram in (a). (c) Gammatonegram of gunshot. (d) Summed localisation response of gunshot in (c).

4. IMPLEMENTATION

4.1 Audio Representation

Three time-frequency representations shall be considered during the evaluation of this system: linear-frequency spectrograms, logarithmic-frequency spectrograms and gammatonegrams. All three representations required parameter tuning so as to reach a trade-off between temporal and frequency granularity, allowing for accurate spectral representations, while maintaining reasonable temporal accuracy.

4.2 Vocabulary

The SIFT algorithm was employed to extract features from the training data for clustering into a vocabulary. Dense SIFT was selected over the traditional SIFT due to the computational expense of the Difference-of-Gaussian extrema selection process. Following extraction, keypoints were normalised before being clustered by means of the K-means algorithm so as to form the model vocabulary. Given the incorporation of spatial data into the image descriptors, a smaller vocabulary suffices, maintaining performance while reducing computation [11]. For this reason, a vocabulary size of 50 was adopted throughout the development and debugging of the system.

4.3 High-Level Representation

Spatial pyramids were found to be the most suitable encoding model for spatial data within descriptors. This strategy incorporates multiple spatial tiles of varying granularity into one encoding so as to sequentially narrow down the region within which a keypoint is situated. A 1x1 grid, encoding no spatial data, is considered at the top of the pyramid (level 0), with each descent resulting in an increase in the number of cells at a rate of 4^x, thus reducing the area covered by each cell; increasing the spatial specificity. Eq. 1 below outlines the manner in which weights are assigned to each of the levels to reflect the spatial precision of each grid.

$$Weight = \frac{1}{2^{current\,level - total\,levels}} \tag{1}$$

A multi-dimensional partitioning data structure, known as a K-D Tree, was then employed for the mapping of features onto their closest vocabulary word, so as to compute the histogram-of-words for each cell. Ultimately, each audio snippet is represented by a vector consisting of histogram-of-words for each of the cells within the spatial pyramid.

4.4 Localisation

The localisation stage is novel in approach, as it is performed prior to classification, as opposed to the majority of pipelines within literature. Test audio files are first transformed into their equivalent visual representation prior to the application of the B-COSFIRE line detection algorithm [12]. This algorithm was tuned to detect vertical lines of 2 pixels in width within the time-frequency representation, with this value achieving the best trade-off for the detection of both impulsive and sustained sounds.

As demonstrated in figures 3(a) and (b), the application of the line detection algorithm resulted in a representation similar to that of the initial gammatonegram, but with a considerable reduction in noise energy. The summed response of the B-COSFIRE algorithm was then taken, with spikes in the plot relating to the detection of suspected events. The audio file is then segmented, with temporal regions of sus-

pected importance undergoing classification.

A suspected event is considered to start at the first non-zero value before a peak, and end at the first zero value after a peak. As may be observed in figures 3(c) and (d), a distribution of the summed response between the start and end of each suspected event was taken. This distribution splits the suspected event into temporal regions, with each region assigned a weighting relative to its normalised localisation score (summed response). Therefore, keypoints at the center of an event signature are considered to be of greater importance than keypoints towards the edges of a signature.

4.5 Classification

SVMs were the classifiers of choice given their flexibility and suitability for application within a computer vision context. Given the inability of SVMs to deal with multi-class problems, and the varying levels of noise existing within the training and testing data, a 28 SVM architecture was proposed. With a total of 4 classification outcomes, and 6 SNR levels (ranging from 5dB to 30dB), 7 SVMs were trained for each event class, with 6 trained upon data of a specific SNR level, and the 7th trained upon data across all SNR levels. The label assigned to the suspected event is dictated by the event class of the SVM achieving the greatest score.

Eq. 2 illustrates the histogram intersection kernel for application within SVM classification. This kernel is found to be highly effective at comparing histogram representations, especially for the classification of colour-based images [13]. Thus, this kernel was deemed to be the most suitable choice for the classification of suspected events of interest.

$$K(a,b) = \sum_{i=1}^{n} min(a_i, b_i) \qquad a_i \geq 0 \qquad b_i \geq 0 \qquad (2)$$

During the implementation of the system, the localisation stage was found to be highly successful at detecting events of importance, but was found to also suspect multiple event-free regions to be of interest, leading to false alarms. Therefore, a neighbourhood reduction process, whereby suspected events within a neighbourhood would be re-evaluated, was proposed as the final stage of the pipeline, so as to reduce the true negative rate of the system. A neighbourhood size was defined so as to perform a sequential search for suspected events across the audio file. Suspected events within the same neighbourhood were compared with respect to localisation score, or alternatively, classification score, with all suspected events except the greatest being deemed noise. A range of neighbourhood sizes were considered, with the optimal being found to be a 3 second diameter. This selection was based upon the assumption that no two events within the dataset could occur within a temporal range of less than 3 seconds.

5. EVALUATION

The MIVIA Audio Events Dataset, on which the proposed system was developed, has been readily split into a training and testing set at a ratio of 70% to 30% respectively by Foggia et al. in [6]. For the purpose of cross-validation, the training set was further split into a training and validation set at a ratio of 80% to 20% respectively.

5.1 Method of Evaluation

Within the setting of the defined system, and as perceived in [2, 6], a true positive (TP) shall be considered to exist whenever any overlap occurs between a suspected event and a ground truth event, with both elements agreeing on an event label. Alternatively, a false positive (FP) shall be considered to occur when a suspected event is erroneously assigned an event label during an event-free interval, while a false negative (FN) occurs when an event of importance is erroneously dismissed as noise.

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+FN} \qquad (3)$$

$$P = \frac{TP}{TP+FP} \qquad FRR = \frac{FN}{TP+FP} \qquad (4)$$

The accuracy of the proposed system shall be measured by considering the true positive rate (TPR), with the false positive rate (FPR) acting as a measure of detection accuracy, allowing for the assessment of the localisation stage. System precision (P) disregards the detection accuracy, allowing for the direct evaluation of the system's classification abilities. Finally, the false rejection rate (FRR) is the ratio measure of events of interest wrongly classified as noise and thus, falsely rejected.

5.2 Experiments

So as to solely focus on the effects of the parameters in question, all other values were standardised during experimentation. Specifically, the Bag-of-Words model with a vocabulary size of 500, no spatial tiling and neighbourhood reduction by localisation score upon a 3 second neighbourhood, were the adopted parameters unless explicitly stated.

5.2.1 Image Representation

Motivation for experimentation upon the different image representations stems from their characteristic differences and the possible effects they may have on the system.

	TPR	FPR	P	FRR
Linear-Frequency Spectrogram	84.65%	6.04%	81.47%	2.11%
Logarithmic-Frequency Spectrogram	91.37%	2.01%	91.44%	2.08%
Gammatonegram	86.71%	4.18%	87.03%	3.90%

Table 1: Table of experimental results for different image representations

Referring to Table 1, in comparison to the other representations, linear-frequency spectrograms are seen to achieve a poor precision in the lower 80 percentile range, illustrating the struggle of classifying such representations. This outcome was anticipated, whereby the linear-scale representation was expected to inadequately express lower frequency bands with respect to their high-pitched counterparts.

Although results improved with the employment of the gammatonegram, the best overall outcome was attained by the logarithmic-scale spectrogram. These results confirm the importance of incorporating logarithmic scales for the classification of events mapped onto the frequency domain. The additional application of the gammatone filter bank to the logarithmic representation was found to hinder results, decreasing classification precision and thus overall accuracy. This experiment thus concludes the suitability of the logarithmic-scale spectrogram representation for audio event segments. All further experiments shall therefore be conducted utilising this representation.

5.2.2 Vocabulary Size and Spatial Information

The spatial encoding procedure employed correlates directly to the feature descriptor size, as does the vocabulary size. Given the relevance between these two impacting system parameters, a joint evaluation was considered suitable.

Vocabulary Size K	Spatial Levels	TPR	FPR	P	FRR
1000	1	91.71%	2.01%	91.78%	2.08%
2000	1	91.59%	4.97%	88.69%	1.70%
200	2	90.76%	4.97%	87.89%	1.70%
400	2	91.97%	2.01%	92.01%	2.05%
50	3	93.03%	2.02%	93.07%	2.05%
100	3	93.26%	2.01%	93.30%	2.05%

Table 2: Table of experimental results for different vocabulary sizes and spatial tiling

Observing the first two rows in Table 2, an increase in the vocabulary size K upon no spatial tiling yielded no noticeable improvements, with a value of 2000 contrarily reducing the TPR, most likely due to over-fitting. The results produced by the 2-level spatial pyramid encodings unexpectedly show a slight decrease in accuracy with K set to 200, and a slight improvement upon the current base-line with a larger vocabulary size. It is thus apparent that a 2-level pyramid does not encode the events' rich visual signatures at a sufficient level of spatial granularity.

Finally, results for the 3-level spatial pyramid implementation exhibit marginal improvements in both precision and general accuracy. This outcome follows previous expectations, as the system is better able to distinguish between different events through the comparison of the spatial location of each of the event keypoints. These parameters, however, result in each event being described by a vector of 21K, lengthening execution time noticeably. Nonetheless, the results were considered satisfactory and were deemed to be computed within a reasonable time-frame.

5.2.3 Neighbourhood Reduction

Neighbourhood reduction is the final stage of the pipeline, and may be employed upon two reduction criteria: localisation score or classification score. The following table provides an accuracy comparison between these different scoring criteria with a neighbourhood size of 3 seconds, so as to verify the relevance of this procedure.

Neighbourhood Reduction	TPR	FPR	P	FRR
Localisation Score	93.26%	2.01%	93.30%	2.05%
Classification Score	90.91%	4.85%	91.02%	4.96%
None	93.53%	12.99%	83.96%	1.78%

Table 3: Table of experimental results for neighbourhood reduction

A marginal reduction in accuracy is observed when classification score is set as the comparator. This outcome is observed due to the struggle of the classifiers to reliably distinguish between events and background noise in suboptimal conditions. The comparison of suspected events by means of localisation score is found to be more suitable, as the events leave a far more distinguishable response across the spectrum, producing a noticeable localisation score.

The most interesting results were achieved in the final test, where no neighbourhood reduction stage was applied. Although the TPR illustrated is the highest accuracy rate achieved during system testing, one must note the great reduction in both the FPR and P rate. These reductions are brought about by the lack of thresholding (filtering) of suspected events during localisation. Considering the proposed application of the implemented system, false alarms are not catastrophic, but must be kept to the utmost minimal. Therefore, within the context of the system, a minimal increase ($\sim 0.25\%$) in FRR, and thus decrease in accuracy, in favour of a reduction in FPR, was considered beneficial.

5.3 Discussion

Following the conduct of the above mentioned experiments, the system was tuned to execute upon the test set employing logarithmic-frequency spectrogram image representation, with the Bag-of-Words model based upon a vocabulary of size 100. A 3-level spatial pyramid encoding was adopted for the description of events, with neighbourhood reduction applied upon a 3 second neighbourhood window, and localisation score acting as the comparator. The complete training set, consisting of 16,726 events, was used to train the system prior to the detection and classification of the test audio files holding 5,376 events of interest.

	TPR	FPR	P	FRR
Proposed System	91.13%	4.35%	89.59%	2.64%
Foggia et al.	86.7%	2.6%	97.7%	10.7%

Table 4: Comparison of system results with those achieved by Foggia et al. in [6]

As one may observe in table 4 above, the proposed pipeline has surpassed Foggia et al.'s system by more than 4% in terms of TPR. This is understandable when comparing the FRR of the two systems, with the proposed pipeline attaining an improved detection rate. This observation in particular confirms the suitability of the proposed localisation stage for such a system. However, while digesting these results, it is vital to understand that while Foggia et al.'s approach provides a solution for real-time audio surveillance, the proposed approach acts in an off-line manner, thus following a far more lenient boundary with respect to computational efficiency.

		Predicted			
		Background Noise	Breaking Glass	Gunshot	Scream
Actual	Background Noise	1814	34	154	46
	Breaking Glass	36	1717	45	2
	Gunshot	55	149	1587	9
	Scream	51	27	103	1595

Table 5: Confusion matrix comparing actual test events with their predicted classification

Table 5 above illustrates the confusion matrix produced by the proposed system, allowing us to better understand the effectiveness of the system at detecting each of the event classes. Breaking glass was by far the easiest event class to detect and classify, most certainly due to the event's distinct signature and spatial layout. The true positive rates for gunshots and screams are slightly inferior, but nevertheless satisfactory. The suitability of event and noise-level specific classifiers, together with the application of the neighbourhood reduction stage is confirmed, with 1,814 audio snippets of suspected interest correctly being dismissed as noise.

Observing the detailed evaluation produced in Table 6, the system's capability of dealing with both impulsive and sustained sounds has been proven, achieving gratifying results within all three classes. As was expected, greater accuracy scores were achieved within optimal conditions, with results

SNR Level	Breaking Glass	Gunshot	Scream	Total	Foggia et al. Total
5dB	93.00%	83.00%	73.65%	83.26%	81.1%
10dB	96.67%	87.33%	89.19%	91.07%	85%
15dB	95.67%	89.33%	94.26%	93.08%	87%
20dB	96.00%	89.00%	94.26%	93.08%	88.4%
25dB	95.33%	90.33%	93.58%	93.08%	88.7%
30dB	95.67%	90.00%	93.92%	93.19%	90%
Total	95.39%	88.17%	89.81%	91.13%	86.7%

Table 6: Breakdown of TPR achieved for each event class at each SNR level. A comparison is made between the results achieved by the proposed system and Foggia et al. in [6]

degrading minimally as the SNR decibel level falls, until a marginal reduction is observed at the 5dB level. Similar trends are observed within the results achieved by Foggia et al., reiterating the challenge of detection and classification posed by highly noisy environments. Observing the final two columns in the table, it is clear that the proposed system out-performs Foggia et al.'s real-time system in all audio conditions, cementing the success of the proposed pipeline.

6. FUTURE WORK

One pitfall of the current implementation is the ineptitude to detect adjoining or overlapping events. Inspiration may be taken from similar applications within literature, such as that of Dennis et al. [8], whereby a Generalised Hough Transform voting system is employed to identify overlapping sound events. The implementation of a separate localisation stage, as opposed to a windowing method, has brought about improvements in efficiency. However, future efforts could be focused on the redesign of a more efficient, but equally effective, classification stage, offering an alternative to the current employment of 28 SVM classifiers.

The procedural steps involved in this pipeline are not domain specific, thus allowing this system to be trained and tweaked for the detection of any set of events. Future work could see the generalisation abilities of the system tested, and the scope expanded to a general event detector, having the system trained to detect dozens of event classes and subclasses, each with their own characteristics and hurdles.

7. CONCLUSION

The ramifications of the selection of image representation schemes has been confirmed, with scaling of the frequency spectrum found to have a direct effect on the results produced. The proposed separation of the localisation and classification stages has proven successful at handling both impulsive and sustained events, which conventional windowing approaches, common within literature, struggle to deal with.

The system has obtained satisfactory results across all SNR levels and event classes, meeting all the aims and objectives laid out from the onset. Through the transformation of audio data onto the visual domain, and the employment of computer vision techniques for the analysis and description of events, a greater robustness to noise was achieved when compared to the direct analysis of low-level audio features. We may therefore confirm the functionality and applicability of computer vision techniques to overcome extensive noise within an automated audio aggression detection system.

8. REFERENCES

[1] Jerry H Ratcliffe, Travis Taniguchi, and Ralph B Taylor. The crime reduction effects of public cctv cameras: a multi-method spatial approach. *Justice Quarterly*, 26(4):746–770, 2009.

[2] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Audio surveillance of roads: a system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288, 2016.

[3] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: a systematic review. *ACM Computing Surveys (CSUR)*, 48(4):52, 2016.

[4] Chloé Clavel, Thibaut Ehrette, and Gaël Richard. Events detection for an audio-based surveillance system. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1306–1309. IEEE, 2005.

[5] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Audio surveillance using a bag of aural words classifier. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 81–86. IEEE, 2013.

[6] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22–28, 2015.

[7] Jonathan Dennis, Huy Dat Tran, and Haizhou Li. Spectrogram image feature for sound event classification in mismatched conditions. *IEEE Signal Processing Letters*, 18(2):130–133, 2011.

[8] Jonathan Dennis, Huy Dat Tran, and Eng Siong Chng. Overlapping sound event recognition using local spectrogram features and the generalised hough transform. *Pattern Recognition Letters*, 34(9):1085–1093, 2013.

[9] Pasquale Foggia, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Cascade classifiers trained on gammatonegrams for reliably detecting audio events. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 50–55. IEEE, 2014.

[10] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[11] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.

[12] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. Trainable cosfire filters for vessel delineation with application to retinal images. *Medical image analysis*, 19(1):46–57, 2015.

[13] Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III–513. IEEE, 2003.