Detecting Violent Crowds using Temporal Analysis of GLCM Texture

Kaelon Lloyd¹ ¹, Paul L. Rosin¹ ¹, David Marshall‡ ² and Simon C. Moore§ ²

¹School of Computer Science, Cardiff University
²Violence and Society Research Group, Applied Clinical Research and Public Health, School of Dentistry, Cardiff University

Abstract

The severity of sustained injury resulting from assault-related violence can be minimized by reducing detection time [10, 28]. However, it has been shown that human operators perform poorly at detecting events found in video footage when presented with simultaneous feeds [30]. We utilize computer vision techniques to develop an automated method of violence detection that can aid a human operator. We observed that violence in city centre environments often occur in crowded areas, resulting in individual actions being occluded by other crowd members. Measures of visual texture have shown to be effective at encoding crowd appearance [8, 23, 31]. Therefore, we propose modelling crowd dynamics using changes in crowd texture. We refer to this approach as Violent Crowd Texture (VCT). Real-world surveillance footage of night time environments and the violent flows [16] dataset were tested using a random forest classifier to evaluate the ability of the VCT method at discriminating between violent and non-violent behaviour. Our method achieves ROC values of 0.98 and 0.91 for the real-world and violent flows data respectively.

1 Introduction

City centre locations around the world are characterized by the presence of surveillance cameras. One typical use of these cameras is to aid law enforcement by allowing operators to actively identify criminal activity. It is estimated that in the United Kingdom alone there are upwards of 1.8 million Closed Circuit Television (CCTV) cameras installed across both public and private sectors, or about one camera for every 35 people, with the average person falling into the viewshed of a camera system at least 68 times a day [7, 12]. An issue with having such a large number of surveillance cameras is that they capture too much data for effective human observation. A study undertaken by Voorthuijsen et al. [30] investigated the human ability to detect scenes of interest from video data when presented with different numbers of simultaneous video feeds. On average, the human ability to detect scenes of interest dropped by 19% when the number of simultaneous feeds was increased from one to four. It is reasonable to assume that the observed drop in event detection ability becomes greater when a single person is subject to the much larger video arrays common in modern observation centres.

Evidence suggests that surveillance systems can reduce the incidence rate of hospital assault-related attendances. Sivaramasringam et al. [28] investigated the relationship between surveillance system installation on violence. It was shown that police recorded violence increased but hospital admissions for assault-related injury fell, an effect that the authors suggest is due to earlier police intervention preventing disorder escalating to the point where serious injury is inflicted on victims of violence. This finding was corroborated by research undertaken by Florence et al. [10] that evaluated data sharing schemes focused on effective strategies for reducing violence. The authors of this study highlight the importance of rapid intervention in reducing injury severity. Although surveillance systems were not the sole focus of this research their usefulness at allowing for effective intervention is acknowledged. A follow up evaluation of the study undertaken by Florence et al. [10] asserts a £1.2 million saving after the application of violence reduction strategies in the year 2007 [11].

There are two practical limitations in using real-world surveillance footage. First the cost of upgrading video capture devices used in surveillance systems across the United Kingdom is high, and accordingly only those cameras that are deemed important are upgraded. It is therefore common for video surveillance systems to be constructed from a mix of both modern and legacy hardware components. The quality of recorded footage from older cameras is typically poor due to hardware limitations and it is common for footage to have low spatial and temporal resolutions. Additionally, outdoor CCTV cameras are subject to natural illumination changes that result in poor contrast when recording footage at night. This makes effective description of content using state-of-the-art techniques difficult. Second, an area in computer vision that is rarely addressed in the literature, violence often occurs in densely populated areas and crowded density can impede classification attempts. Footage of dense populations in urban environments tends to depict moving, self-occluding crowds. It can be difficult to generate a meaningful description of individuals and their actions as the visual consistency of recognizable shapes fluctuates greatly between frames due to the high levels of occlusion. In recognizing the potential value of surveillance systems in reducing assault-related injury this paper describes a novel solution to the previously discussed limitations, and therefore provides opportunities for the classification of live surveillance feeds to aid operators’ early ascertainment of violence and therefore their capacity to direct resources that stop disorder escalating.

Our novel solution builds upon the idea of describing crowded scenes using visual texture. Texture is well suited for describing the seemingly unstructured patterns that result

¹ Lloyd@cardiff.ac.uk
² Paul.Rosin@cs.cardiff.ac.uk
³ MarshallAD@cardiff.ac.uk
⁴ mooresc2@cardiff.ac.uk

Evidence suggests that surveillance systems can reduce the incidence rate of hospital assault-related attendances. Sivaramasringam et al. [28] investigated the relationship between surveillance system installation on violence. It was shown that police recorded violence increased but hospital admissions for assault-related injury fell, an effect that the authors suggest is due to earlier police intervention preventing disorder escalating to the point where serious injury is inflicted on victims of violence. This finding was corroborated by research undertaken by Florence et al. [10] that evaluated data sharing schemes focused on effective strategies for reducing violence. The authors of this study highlight the importance of rapid intervention in reducing injury severity. Although surveillance systems were not the sole focus of this research their usefulness at allowing for effective intervention is acknowledged. A follow up evaluation of the study undertaken by Florence et al. [10] asserts a £1.2 million saving after the application of violence reduction strategies in the year 2007 [11].

There are two practical limitations in using real-world surveillance footage. First the cost of upgrading video capture devices used in surveillance systems across the United Kingdom is high, and accordingly only those cameras that are deemed important are upgraded. It is therefore common for video surveillance systems to be constructed from a mix of both modern and legacy hardware components. The quality of recorded footage from older cameras is typically poor due to hardware limitations and it is common for footage to have low spatial and temporal resolutions. Additionally, outdoor CCTV cameras are subject to natural illumination changes that result in poor contrast when recording footage at night. This makes effective description of content using state-of-the-art techniques difficult. Second, an area in computer vision that is rarely addressed in the literature, violence often occurs in densely populated areas and crowded density can impede classification attempts. Footage of dense populations in urban environments tends to depict moving, self-occluding crowds. It can be difficult to generate a meaningful description of individuals and their actions as the visual consistency of recognizable shapes fluctuates greatly between frames due to the high levels of occlusion. In recognizing the potential value of surveillance systems in reducing assault-related injury this paper describes a novel solution to the previously discussed limitations, and therefore provides opportunities for the classification of live surveillance feeds to aid operators’ early ascertainment of violence and therefore their capacity to direct resources that stop disorder escalating.

Our novel solution builds upon the idea of describing crowded scenes using visual texture. Texture is well suited for describing the seemingly unstructured patterns that result

¹ Lloyd@cardiff.ac.uk
² Paul.Rosin@cs.cardiff.ac.uk
³ MarshallAD@cardiff.ac.uk
⁴ mooresc2@cardiff.ac.uk
from the mass occlusions caused by crowding [23]. Generally speaking, the appearance of violent crowds undergoes greater amounts of change in a shorter time span compared to non-violent crowds. We therefore suggest a description based on encoding the change in crowd appearance over time. This is accomplished by computing texture features on a per-frame basis for a sequence of video frames and summarising how the texture features evolve.

We show that our proposed method, Violent Crowd Texture (VCT), generates a scene description that can be used to discriminate between violent and non-violent scenes available in the Violent Flows dataset. We perform tests using real-life data captured from CCTV cameras installed in city centres. Additionally, we conduct tests on the UMN unusual crowd and Web abnormality datasets to determine how effective the VCT descriptor is at classifying crowd dynamics that are typical of city centre locations but fall outside of the violent and non-violent class labels. Our method attains receiver operating characteristic (ROC) performance values of 0.98, 0.91, 0.97 and 0.93 for the real-world, Violent Flows, UMN and Web abnormality datasets respectively. It is demonstrated that the VCT representation offers consistently high performance across many different types of data. We compared our results with those obtained from four other state-of-the-art methods; these are STIP [20], LBP-TOP [33], Violent Flows [16] and MBP [3]. We found that our method outperforms STIP and MBP in all cases. LBP-TOP and Violent Flows offer comparable results on specific datasets but unlike results attained using VCT, they fail to attain consistently high performance across all four tested datasets.

2 Related Work

Research in the area of action recognition has advanced significantly in the past decade. Earlier research was conducted on simple actions performed by a single individual. Common datasets for testing simple action recognition ability are the KTH [27] and Weizmann [4] datasets. Impressive results using a wide range of methods that exploit many different visual aspects such as texture, colour, shape and optical flow in unique ways have been attained using these datasets. The high performance achieved on single action datasets led to the construction of more complex datasets such as the widely explored UCF50 [26] which contains footage of many different sporting activities. Other challenging datasets such as Hollywood2 [24] contain more complex human actions and currently prove difficult to solve even with state-of-the-art techniques.

Violent detection in city environments has not been widely investigated within the computer vision community. Efforts have been made in the realm of violence detection but most research in this area is conducted on data that is not indicative of footage captured by real-world surveillance cameras, for example Lin [21] and Chen [6] used visual and audio cues to identify violent scenes in Hollywood films. Early work by Datta [9] investigated one-on-one violence, however the method they proposed requires the successful interpretation of human silhouettes and bounding boxes, a requirement not easily satisfied when dealing with crowded street environments. The Violent Flows (ViF) method was proposed by Hassner et al. [16] to identify violent crowds in densely populated areas using changes in optical flow magnitude. Gao et al. [13] states that the ViF descriptor does not capture potentially important changes in orientation and so introduced a variant of the ViF descriptor (OViF) that utilises both orientation and magnitude of optical flow. It is shown that ViF offers greater classification ability on crowded data when compared to OViF. The work by Hassner et al. offers a robust point of comparison for the method proposed in this paper.

Many state-of-the-art methods adopt a local description approach that relies on salient point detectors to identify regions of highly descriptive information. Space Time Interest Points (STIP) is one popular and widely used method in which the Harris corner detector is extended into the third, temporal dimension [19]. STIP points have been adopted by many researchers to identify salient regions for further description using techniques such as Histogram of Oriented Flows (HOF) + Histogram of Oriented Gradients (HOG) [20]. Volume based Local Binary Pattern [2] or the spatio-temporal extension of HOG known as HOG3D [18]. These method prove effective at individual action recognition with classification accuracy values of 91.1%, 90.55% and 92.6% respectively when applied to the KTH dataset. Although salient point detection can offer good performance Kadir and Brady [17] note that response functions, by the nature of their design, introduce some geometric or morphological constraints that may lead to the inadequate or the inappropriate detection of salient points. Consequently, generating a useful salient point detector for crowded scenes found in city centre locations would be difficult due to the large variation of the seemingly stochastic patterns that result from self-occluding crowds.

Global methods operate by describing the entire frame or frame parts, and are suitable for providing context to an entire scene where individual actions cannot be easily realised. It is for this reason that Hassner et al. [16] used a global representation in the description of violent crowds. Ali & Shah [1] segment crowded motion by treating a moving crowd as a global entity and modelling the perceived motion as a fluid flow rather than tracking local points. Global representations are intuitively more reliable than the local alternative as salient point detection is not guaranteed.

Early research by Marana [23] formulated the crowd density estimation problem as a global measure of visual texture. Marana showed that sparse and dense crowds hold notably different textural compositions. Researchers [8, 23, 31] have used the Grey Level Co-Occurrence Matrix (GLCM) approach to crowd description and have shown that Haralick’s GLCM features can be used to successfully determine the density of a crowd. The implication being that texture can provide a meaningful description of the visual appearance crowds.

3 Violent Crowd Texture Method Overview

Our proposed method builds upon Haralick texture features [15] which describe visual texture using statistics derived from grey level intensities. We suggest computing Haralick features for each frame in a sequence and describing how these features evolve over time using simple summary measures to provide a succinct and powerful descriptor of crowd dynamics. Haralick texture features are extracted from a grey level co-occurrence matrix (GLCM). A GLCM is generated by counting the co-occurring grey level intensity values found in an image given a linear spatial relationship between two pixels. The spatial relationship is defined by a parameter pair \((\theta, d)\) where \(\theta\)
is the orientation and \( d \) is the distance between two pixels. It is common to define a set of parameter pairs \((\theta, d)\) and then combine the resulting GLCMs by summing them together. One strong reason for doing so is to provide rotational invariance by using a set of orientation parameters, typically in 8 orientations, spaced \( \pi/4 \) radians apart. The number of grey level values \( N_g \) represents the number of unique intensity values present in an image. It is common to scale an image from \([0, 255]\) to \([0, N_g]\) before computing a GLCM matrix. Computing all 14 of Haralicks outlined methods and attaining a real-time system is not feasible. We performed feature selection to identify four features that offered good results under a real-time constraint. The measures we adopt are Energy, Contrast, Homogeneity and Correlation as defined by Haralick [13].

We compute the aforementioned texture features for each frame in a video sequence and describe a scene based on the change in feature values over time, each measure is represented as a separate distribution \( x \). Our descriptor adopts the dispersion measures mean and standard deviation of \( x \) which provide a general understanding of the appearance and amount of variability that is found within an image sequence. These measures are descriptively broad and remove the finer temporal details that may exist. We therefore introduce the skewness central moment and uniformity measures to provide detail.

**Skewness** indicates the asymmetry found in a distribution and can be used to deduce whether a distribution is showing a general increase or decrease in value over time.

\[
Skewness = \frac{E((x - \mu)^3)}{\sigma^3}
\]

**Uniformity** provides a measure of similarity between adjacent samples with respect to time by computing the absolute difference between adjacent values. This formula output limits are \([0, 1]\).

\[
Uniformity = \frac{\sum_{t=1}^{T-1}|x_t - x_{t+1}|}{\max x \times (T - 1)}
\]

It was observed that different spatial regions depicted different behaviour, therefore we spatially split each video into \( M \times N \) non-overlapping sub-regions and apply the aforementioned method to each. The final descriptor is a concatenation of the feature vectors extracted from these regions. In the case of surveillance footage, failure to remove background information may lead to the description of landmarks as opposed to crowd dynamics. It is observed that static regions occurring across multiple frames in the sequence contribute the same information when forming a GLCM. We therefore perform frame subtraction at the GLCM level of the processing pipeline. This approach comes at a near negligible computational cost and offers robustness to minor translational camera motion due to the spatially unconstrained nature co-occurrence matrices. The spatial constraint of the GLCM subtraction approach is defined by the \( M \times N \) grid structure previously mentioned.

### 3.1 Alternative Form using Optical Flow

If robust optical flow fields can be computed from the source data then benefits may be seen by substituting grey level intensities with optical flow information. The process involves computing dense optical flow fields using the Lucas-Kanade [22] approach. The next step is to derive motion magnitude and orientation at each point using Equations 3 and 4. Placing these values into separate arrays and scaling their values allows this data to be used with the method outlined in the previous section. Intuitively, the main diagonal in the GLCM will describe the quantity of uniform motion while the off-diagonal values will provide context of motions and their neighbouring reactive forces. Throughout this paper we refer to this variant as T+OF (Texture + Optical Flow).

\[
m_{x,y} = \sqrt{u_{x,y}^2 + v_{x,y}^2}
\]

\[
\theta_{x,y} = \tan^{-1}\left(\frac{v_{x,y}}{u_{x,y}}\right)
\]

### 4 Data

The goal of our research is to provide a computational method that can aid CCTV operators at detecting violent scenes and so we conduct tests using the Violent Flows dataset and real-world CCTV data captured in city centre environments. Additionally, detecting abnormal crowd behaviour can be beneficial as abnormal events may lead to violent situations. Therefore, we conduct experiments on two crowd abnormality datasets that will give insight into the generality of the VCT method at describing crowd behaviour.

#### 4.1 City Centre Environment (Real-World)

Data

We obtained real-life surveillance footage from a local police force that showed either violent or non-violent behaviour within city centre locations. Experiments performed on this data will provide a realistic understanding of the applicability of each tested method in a real-world scenario. We obtained 13 samples of violent behaviour and 63 samples of general behaviour. The violent scenes can be separated into two distinct classes of high and low based on the participant population. Only 4 of the 13 samples can be considered to have a large number of participants. Due to the low number of samples we perform 4-fold cross validation. Video resolutions range between 320 \( \times \) 240 and 640 \( \times \) 480 and all videos were recorded at a de-interlaced frame-rate of six frames per second. Surveillance cameras are typically placed at a high altitude in order to maximize viewshed. The elevated height makes a camera more exposed to high winds which causes the camera to shake and capture spatially unstable footage; this can cause issues when trying to identify corresponding features between adjacent frames. All real-world footage is stabilized before subsequent analysis using the state-of-the-art method stabilization method proposed by Grundman et al. [14].

#### 4.2 Violent Flows

The Violent Flows dataset [16] was created for the sole purpose of evaluating crowd violence classification methods; it is a relatively new dataset and is not widely tested. There are 123 instances of both violent and non-violent data samples available from footage uploaded to video media hosting websites. The violent footage contains many samples that are visually similar to those found in real-world data and it is therefore suitable for evaluating the violence classification properties of the VCT representation.
4.3 UCF Web Crowd Abnormality

The UCF web crowd abnormality dataset consists of 20 videos depicting either normal or abnormal crowd behaviour [25]. Abnormal data is classified as panic, clash or fight scenarios. Normal samples can be described as showing either crowds walking in an urban environment or pedestrians running in a marathon. The dataset has been obtained from various media hosting websites and of the 20 available sequences, 12 are normal and 8 abnormal. Footage is recorded at 24 frames per second with a resolution of 640 × 480. Image examples of the web crowd abnormality dataset can be seen in Figure 2.

4.4 UMN Crowd Abnormality

The UMN unusual crowd activity dataset [25] is a synthetic set that depicts sparsely populated areas. Initially, normal crowd activity is observed until a specified point in time where behaviour rapidly evolves into an escape scenario where each individual runs out of camera view to simulate panic. The dataset is comprised of 11 separate video samples that start by depicting normal behaviour before changing to abnormal. The panic scenario is filmed in three different locations, one indoors and two outdoors. All footage is recorded at a frame rate of 30 frames per second at a resolution of 640 × 480 using a static camera.

5 Experiments

A classification label is generated for each video frame in order to provide a continuous activity feed usable in CCTV observation scenarios. This is achieved by classifying a description vector computed using the previous n frames in sequence where n is equal to the number of frames per second. For the generation of the grey level co-occurrence matrix used in the VCT descriptor we set \( N_g = 32 \). The parameters \((\theta, d)\) are assigned as \((0, 1)\), see section 6.1 for an explanation regarding the choice of these parameters. The T+OF approach (section 3.1) utilizes GLCM matrices for both the magnitude and orientation of optical flow flow. A set of distance values are selected \( d = \{1, 2, 4, 8\} \), \( N_g = 20 \) for the flow magnitude data and \( N_g = 8 \) for the optical flow orientation data; the GLCM orientation parameter \( \theta \) is assigned as a set of 8 orientations, spaced \( \pi/4 \) radians apart. \( M \) and \( N \), which specify frame sub-division used to encode spatial information are assigned the value of 4, section 6.2 discusses the effects of using different values for \( M \) and \( N \).

All experiments were conducted using k-fold cross validation where data is split into \( k \) partitions with \( k - 1 \) partitions being used for training a random forest classifier [5]. The remaining partition is used for testing; the random forest is composed of 25 trees. We perform each experiment 100 times and report the average result to reduce any variability introduced by random sampling during cross validation. As stated previously, we extract features such that each frame in a sequence is represented by a single vector, we do not allow features extracted from a single source video to be placed in both training and testing partitions at the same time as features extracted from any single video are likely to belong to the same distribution and may lead to over-fitting. We present results using ROC curves, a common way to summarise these curves is to report area under the curve. Area under ROC dictates the discrimination performance between binary classes, a value of 1 indicates perfect discrimination. VCT was implemented using an amalgamation of Matlab 2014a and C++. The generation of grey level co-occurrence matrices was too time consuming when implemented in Matlab so we migrated that module to
C++ in order to attain real-time processing on an Intel i7-4790 at 3.6GHz.

We cannot visually validate the GLCM background subtraction approach, as transforming a GLCM back to an image is impossible once subtraction is performed. We therefore perform additional experiments where we substitute the GLCM based background subtraction with alternative methods to provide a comparative baseline for evaluation. The methods we adopted were adaptive differencing [22] and a Gaussian mixture model approach [29]; these methods were chosen for their computational efficiency and therefore suitability for real-time systems.

We tested state-of-the-art methods as listed in Table 1 and to ensure consistency we performed all tests using the same testing strategy outlined prior. LBP-TOP utilizes the same $M \times N$ grid as VCT where $M = N = 4$. Where possible we obtained code from the author’s website. The methods listed in Table 1 were chosen as they perform well according to their respective literature.

Referring to Table 1 and Figure 4, it can be seen that VCT outperforms all other approaches when applied on the real-world dataset. Table 2 displays the accuracy of each method at classifying scenes of violence (true positive) and scenes of non-violence (true negative). All methods show excellent ability at correctly identifying scenes of non-violence; however, only VCT attains a satisfactory classification score when detecting scenes of violence. The real-world data is characterized by a low temporal sample rate; objects in frame show a greater spatial displacement as they move across adjacent frames compared to footage recorded at the typical 25 or 30 frames per second. This low frame rate effects methods that rely on the explicit description of local changes; Lucas-Kanade optical flow (used in T+OF and Violent Flows) assumes that a pixel in the next frame exists within a local neighbourhood of a pixel in the prior frame, the large displacements caused by a low frame rate reduces the probability that a pixel will exist within the neighbourhood region resulting in incorrect flow generation. The VCT descriptor is less affected by this as it does not explicitly describe local correspondences between adjacent frames.

Table 1 shows that the model based background subtraction approaches GMM and Adaptive Difference provide poorer performance compared to the approach described in Section 3 across all datasets. This is prevalent when testing using the Violent Flows dataset, see Figure 2. To explain this we must describe the data. Visually observing the Violent Flows dataset we find multiple instances of footage recorded using handheld cameras. Hand operated devices typically depict unstable footage as a result of shaking and sweeping motions caused by the human operator attempting to follow a scene of interest. Additionally, model-based approaches use appearance history consisting of the past few frames of video to generate and update a model that can be used to identify background information. A model-based approach updates gradually meaning that old information dissipates over time, therefore any erroneous additions to the model will linger. Unstable footage is more prone to introducing errors into the model which will remain over a period of time. The alternative subtraction approach as described in Section 3 does not rely on an appearance history but rather on instantaneous difference, therefore any errors that result in this process are not propagated. Additionally the reduced spatial constraint imposed by the GLCM process makes the subtraction approach robust to small camera motion. Regarding the Violent Flows dataset, we see that LBP-TOP obtains the best score on the Violent Flows dataset but comes with the caveat that it cannot be executed in real-time on the test machine.

Table 1 and Figure 5 demonstrate that T+OF outperforms all other methods when applied to the UMN dataset with comparable results found with the LBP-TOP and VCT with either GMM or GLCM subtraction. Figure 5 shows that VCT is best suited for tackling the challenging Web abnormality dataset.

6 Parameter Effects

In this section we will look at how each parameter affects classification performance.

6.1 Pixel pair relationship

As stated in section 3, a grey level co-occurrence matrix is generated by counting pixel pair occurrences given a relationship defined by parameters $(\theta, d)$. We evaluate the effects of these parameters by performing multiple experiments with a range of values. Our experiments use each combination that can be composed from one of three orientation configurations and one of five different distance values, this provides 15 different ex-
Table 1: Area under ROC curves for each dataset

<table>
<thead>
<tr>
<th>Method (Background Removal)</th>
<th>Real World</th>
<th>Violent Flows</th>
<th>UMN</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed (Adaptive Difference)</td>
<td>0.9712</td>
<td>0.8367</td>
<td>0.9489</td>
<td>0.8913</td>
</tr>
<tr>
<td>Proposed (GMM)</td>
<td>0.9501</td>
<td>0.8462</td>
<td>0.9719</td>
<td>0.8866</td>
</tr>
<tr>
<td>Proposed (Subtraction)</td>
<td><strong>0.9838</strong></td>
<td>0.9134</td>
<td>0.9732</td>
<td><strong>0.9287</strong></td>
</tr>
<tr>
<td>Proposed (T+OF)</td>
<td>0.9502</td>
<td>0.8611</td>
<td><strong>0.9845</strong></td>
<td>0.5837</td>
</tr>
<tr>
<td>LBP-TOP</td>
<td>0.8998</td>
<td><strong>0.9386</strong></td>
<td>0.9724</td>
<td>0.6882</td>
</tr>
<tr>
<td>STIP HOG+HOF</td>
<td>0.6136</td>
<td>0.6723</td>
<td>0.6816</td>
<td>0.7110</td>
</tr>
<tr>
<td>Violent Flows</td>
<td>0.7443</td>
<td>0.8367</td>
<td>0.8212</td>
<td>0.5929</td>
</tr>
<tr>
<td>MBP</td>
<td>0.8430</td>
<td>0.6523</td>
<td>0.6518</td>
<td>0.6555</td>
</tr>
</tbody>
</table>

Table 2: Real-world data classification accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>TN</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed (Adaptive Difference)</td>
<td>71.58</td>
<td>97.82</td>
<td>84.70</td>
</tr>
<tr>
<td>Proposed (GMM)</td>
<td>60.38</td>
<td>97.34</td>
<td>79.10</td>
</tr>
<tr>
<td>Proposed (Subtraction)</td>
<td><strong>83.94</strong></td>
<td>97.89</td>
<td><strong>90.88</strong></td>
</tr>
<tr>
<td>T+OF</td>
<td>70.00</td>
<td>97.34</td>
<td>83.67</td>
</tr>
<tr>
<td>LBP-TOP</td>
<td>58.13</td>
<td><strong>99.68</strong></td>
<td>78.91</td>
</tr>
<tr>
<td>STIP HOG+HOF</td>
<td>46.51</td>
<td>96.42</td>
<td>71.47</td>
</tr>
<tr>
<td>Violent Flows</td>
<td>56.29</td>
<td>86.12</td>
<td>71.20</td>
</tr>
<tr>
<td>MBP</td>
<td>56.49</td>
<td>89.27</td>
<td>72.88</td>
</tr>
</tbody>
</table>

Experiments. The first orientation configuration is a set of 8 orientation values spaced $\pi/4$ radians apart, the second set contains 4 orientation values spaced $\pi/2$ radians apart. The final orientation set contains a single value of 0. The five distance values are a sequence of integers that double in size starting from 1. The results of this experiment are shown in figure 8.

With the exception of the UMN dataset we see a common trend. When looking at the effects of the distance parameter we see a decline in classification ability as the parameter $d$ increases. This trend is more prevalent in results obtained from the web abnormality and real world datasets with an average difference between ROC score where $d = 1$ and $d = 16$ being 0.13 and 0.05 respectively. We hypothesise that using a large distance value creates a situation where the texture describes the relationship between distant objects in a scene that are not directly related, and therefore provides a description that is less meaningful. By selecting $d = 1$ we are more likely to model close interactions between pedestrians regardless of captured scale. The sparse nature of the crowds that appear in the UMN dataset allows for less occlusions and so larger distance values are less likely to create a relationship between two uncorrelated regions in frame. Once again, with the exception of the UMN dataset it can be observed that using a single orientation value provides the best performance on all datasets, and that the difference between the 4 and 8 orientation configurations is negligible. As stated in section 3 using multiple orientations together offers rotational invariance depending on the orientation set used. This is not apparent from the results taken from this experiment. To show that the stated invariance exists, a second test was conducted where two sets of features were extracted. The first set of features is extracted as stated in section 3. The second set of features is extracted using the same approach but each frame from the sequence is rotated randomly by either 90, 180 or 270 degrees. To demonstrate that rotational invariance is obtained, we assign grid separation values $M = N = 1$ so that spatial localization is removed. We find that when applied on the UCF dataset, using 4 orientations spaced $\pi/2$ radians apart
resulted in a ROC score of 0.9027, this is a difference in ROC of 0.0795 (9.21%) when compared against the description using a single orientation ($\theta = 0$) which achieved 0.8232.

### 6.2 Frame Split

In section [6], we suggest splitting the frame into sub-regions to introduce spatial information into the descriptor. In this section we will explore the performance obtained when using different grid sizes. This experiment uses exponentially increasing values for $M$ and $N$ and constrains the tests such that $M = N$. The results are presented in figure [7]. When applied to the UMN panic dataset we find that using $M = N = 1$ results in the worst performance, and that a notable boost in observed $\theta = 0$ which achieved 0.8232.

Figure 8: Graph shows the effects of pixel pair relationship parameters: (a) UCF Web Abnormality, (b) Real World CCTV, (c) UMN panic, d) Violent Flows

Figure 9: Shows the effects of using different values for $M$ and $N$ where $M = N$

![Figure 9](image)

Figure 10: Shows the effects of using different window sizes $n$

this occurs but it is hypothesised that the summary statistics used in VCT fail to describe prevalent changes that contribute a minority of the information contained in the window. Although each dataset has its preferential window length, it is important to note that short window sizes still offer reasonably good performance across all datasets. When transitioning from normal to abnormal behaviour, the amount of time required for the majority of the feature vector to be composed of information from abnormal behaviour will be greater the larger the observation window. Assuming that class transitions are not represented by the descriptor, the worst case for classifying abnormal behaviour will see a delay of at most $n$ frames. Therefore shorter observation windows are more appropriate for use in a real-time system as it will allows for more instantaneous updates.

### 6.3 Window Length

In section [7] we state that the past $n$ frames in sequence are used to form our descriptor for a frame at a given time. In this experiment we will assign different values to $n$. We perform classification using descriptors formed using the following values of $n$: 2, 6, 12, 24, 32, 64 and 128. Looking at the ROC results shown in figure [10] we find that no single window length performs better than the rest across all datasets and that the results are generally stable over values of $n$. The exception is the UCF web abnormality dataset sees a vast decline in classification ability at larger window sizes. It is unknown exactly why

![Figure 10](image)

Figure 10: Shows the effects of using different window sizes $n$

7 Conclusion and Future Work

In this paper we have tested several state-of-the-art action recognition techniques to determine a method that best describes violent situations that occur in city centre environments. We propose using GLCM texture features that are typically
used in crowd density estimation and applied temporal encoding to create an effective method that describes crowd dynamics. We have demonstrated that the proposed method is highly effective at discriminating between scenes of violence and non-violence. We have also shown that the proposed method can be used to discriminate between standard behaviour and other abnormal activities that may occur within city centre locations. In section 5 we set the number of frames n that constitute a sample for classification to the source video frame rate, however further research can be conducted to determine a method of adaptively choosing the optimal value that will deliver the best performance. We would also like to apply computer vision based violence detection systems in the real-world to evaluate their ability at reducing the impact of violence related injuries by assisting CCTV observation personnel.

References


