Violent Behaviour Detection using Local Trajectory Response

K. Lloyd*, P.L. Rosin*, A.D. Marshall*, S.C. Moore+

*School of Computer Science, Cardiff University, United Kingdom. LloydK1@cardiff.ac.uk
+Violence and Society Research Group, School of Dentistry, Cardiff University, United Kingdom.

Keywords: Violence, Trajectories, Detection, Surveillance.

Abstract

Surveillance systems in the United Kingdom are prominent, and the number of installed cameras is estimated to be around 1.8 million. It is common for a single person to watch multiple live video feeds when conducting active surveillance, and past research has shown that a person's effectiveness at successfully identifying an event of interest diminishes the more monitors they must observe. We propose using computer vision techniques to produce a system that can accurately identify scenes of violent behaviour. In this paper we outline three measures of motion trajectory that when combined produce a response map that highlights regions within frames that contain behaviour typical of violence based on local information. Our proposed method demonstrates state-of-the-art classification ability when given the task of distinguishing between violent and non-violent behaviour across a wide variety of violent data, including real-world surveillance footage obtained from local police organisations.

1 Introduction

1.1 Motivation

It is common that major city centre locations are under the constant gaze of surveillance cameras. 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. To provide some perspective, it is estimated that one camera for every 35 people is operational in Britain, and that the average person falls into the viewshed of a camera system at least 68 times a day [5, 16].

Sivarajasingam *et al.* [18] investigated the ties between the installation of surveillance systems and their effects on assault related injury statistics. They suggest that the effectiveness of surveillance systems lies in the detection process rather than the prevention of violent situations. Specifically they looked at the correlation between CCTV installation and the quantity of assault related injury treated at emergency departments. A reduction in hospital admittance was observed and the authors hypothesised that active camera surveillance allows for quicker detection of violent behaviour that leads to quicker intervention, which in turn reduces the seriousness of sustained injury. This idea was corroborated by research undertaken by Florence *et al.* [3] which saw the implementation of a data sharing scheme that focused on devising effective strategies for the reduction of violence using statistics obtained from relevant areas such as hospital emergency departments.

In a study undertaken by Voorthuijsen *et al.* [19] a number of participants sat through two hours of recorded surveillance footage while noting any incidents depicted; the results showed that when presented with four video feeds at once, the detection rate drops from 91% down to 72% compared to monitoring a single video feed. A standard CCTV observation centre can have hundreds of simultaneous feeds shared between three or four people [6, 15]. A further decrease in human observation ability is expected when extrapolating the results found by Voorthuijsen *et al.*

In this paper we present an approach for active violence surveillance using computer vision techniques to identify and describe potentially violent regions in frames. We present a scale-invariant detection scheme that focuses on identifying violent behaviour characterised by high acceleration interactions involving non-linear motions. We evaluate the effectiveness of our approach at classifying violent behaviour using four datasets, two of which are composed of real-world footage of city centre environments from two cities located in the United Kingdom.

1.2 Related Work

Datta [1] and Deniz [2] produced methods that detect violence by identifying high motion acceleration, a property expected to belong to violent behaviour. Deniz notes that high acceleration often manifests itself as a visual motion blur which can be measured by identifying the shape of an ellipse in a Radon transformed power spectrum composed using two consecutive images. Datta takes a more structured approach to solve the task of measuring person-on-person violence by first determining a person's silhouette, and subsequently their head, for tracking. The third derivative of motion, known as jerk, is then incorporated in the composition of the Acceleration Measure Vector to describe violent behaviour. The drawback of this work is that it assumes a person's body is visible and trackable which in a city centre environment is not feasible due to occlusions caused by pedestrians in populated areas.

In contrast to person-on-person violence, Hassner *et al.* [8] looked at differentiating between violent and non-violent crowds. The authors introduced the Violent Flows dataset alongside the Violent Flows (ViF) feature vector that measures

the average magnitude of dominant motions over time. Gao *et al.* [4] state that ViF does not capture changes in orientation and demonstrated that an orientation focused variant of ViF, known as Oriented ViF (OViF), increased classification ability when combined with ViF features.

Nievas *et al.* [14] extended the SIFT descriptor to work on optical flow data and created Motion Scale Invariant Feature Transform (MoSIFT). They used a combination of MoSIFT and SIFT features to classify between violent and non-violent scenes that occur in ice hockey. Xu *et al.* [20] used the work by Nievas and applied a kernel density estimation process to select the most important features before applying a sparse coding scheme. This allowed MoSIFT features to achieve excellent classification on person-on-person violence as well as on crowded data. Gracia *et al.* [7] argue that approaches such as these, although impressive, are too computationally costly to be practically implemented in the real-world, so they propose a more efficient method of violence detection. They perform adjacent frame differencing and apply a fixed threshold to extract the biggest blobs which are then described using measures of distance and compactness.

Riberio *et al.* [17] introduce the Rotation Invariant Motion Coherence RIMOC feature that is based on the eigen-values of second order statistics extracted from a Histogram of Oriented Flows. A multi-scale structure is used to model spatio-temporal configurations of features. The authors assume that violent behaviour is unstructured and aim to distinguish this difference by analysing the likelihood of a feature belonging to a model of normality.

Our work is more akin to that of MoSIFT or STIP [10] but with a focus on using trajectory dynamics to identify spatio-temporal regions that exhibit non-linear, high acceleration interactions that are typical of violent behaviour. Our proposed approach is designed to function well on real-world surveillance data and is rotation and scale-invariant.

2 Proposed Method

We propose a linear combination of three measures of motion trajectory to produce a response map that highlights spatio-temporal regions which are suggestive of violent behaviour. Motion trajectories are computed using a particle advection process that is widely used in pedestrian analysis due to its robustness to minor occlusions [12, 13, 22]. We perform local response normalization across a range of local neighbourhood sizes and apply space-scale non-maximum suppression to achieve a multi-scale system.

2.1 Particle Advection

The process of advection can be defined as the movement of an object through a medium guided by an underlying flow field; this process can be envisioned as a leaf in a river flowing downstream guided by the local flow of the water. In the context of our work, particle advection is performed by first generating a uniform set of particles that overlay the initial frame of a spatio-temporal volume. Each particle is advected using a Gaussian average of local optical flow vectors computed between successive frames. We initialize the particle advection grid to be equal to the dimensions of the video being analysed and advect for τ frames in order to produce a dense trajectory set T.

2.2 Inverse Laminar Flow

It has been observed that pedestrians walking through an environment tend to exhibit laminar flow as they walk towards their destination providing their pathway is not obstructed [9, 21]. Pedestrians are not the only entities that exist in a city centre environment that exhibit this behaviour, with a vehicle being the most prominent example. We also observed that participants of a violent situation were often unstable in their movements as they attempted to perform violent gestures towards another person, therefore we suggest incorporating an inverse laminar flow measure as an indicator of potential violent behaviour. For each trajectory T we compute the total distance travelled T_{dist} over τ frames, and the displacement T_{disp} defined as the absolute difference between the trajectory's initial starting position and its final position after τ frames. These two values are combined to form the inverse laminar flow response R_{ilf} shown in Equation 1.

$$R_{ilf} = 1 - (T_{disp}/T_{dist}) \tag{1}$$

2.3 Acceleration Response

It has been observed that violent acts tend to show a greater increase in velocity when compared to normal behaviour. This was also one of the key principles behind the Maximum Warping Energy method of violence recognition [11]. Given a set of trajectories we can determine the acceleration response R_a for each trajectory T by dividing the maximum acceleration of a single trajectory T by the maximum acceleration of all trajectories.

2.4 Motion Convergence

The location of an interaction between multiple entities can be identified by determining the point where they converge if they were to continue with their respective motion. For example, in a real-life scenario this can manifest itself as two people approaching one another, or to use a more extreme example, a fist moving towards a person's body. We propose identifying local trajectory convergence to describe whether or not trajectories interact. This is achieved by generating a 2D-histogram using the (x, y) position reported at the end of each trajectory; the histogram's dimensions are equal to the spatial dimensions of the video being analysed. A value greater than one in the histogram indicates that multiple trajectories have converged as they occupy the same position in space. We subtract a value of one from each position in the histogram to remove all non-convergent points, followed by the application a Gaussian blur to smooth results. To form the response map R_c we assign a value at the starting position of each trajectory with a value equal to the magnitude of the histogram bin that the trajectory contributes to.

2.5 Response Map

We combine the responses outlined in previous sections as a linear combination (Equation 2).

$$R = (w_1 R_{ilf} + w_2 R_c + (1 - w_1 - w_2) R_a) \quad (2)$$

Often, surveillance cameras are placed to maximize viewshed which can result in a captured scene that contains actions that occur at different distances from the camera due to perspective; this will affect our responses. For example, we will often observe high acceleration response R_a at regions in frame that are closer to the camera. To alleviate this issue we propose normalizing the response using local information defined by a local neighbourhood region $N \times N$. For each point in a response map we subtract the local Gaussian average and normalize by the local maximum within the same region. This will alter the response such that points that are both locally maximal and locally contrasting are given a greater value. This also has the effect of bounding the response within $[-1, 1]$, however we truncate all negative responses to zero.

For each point in the response map R, we apply the aforementioned process using a range of neighbourhood sizes to create a response for each scale, we then subtract adjacent scale responses to form a collection of response scale gradients. We compute the mean gradient μ and starting at the smallest scale, moving towards the largest, we determine if the gradient between two adjacent scale responses is greater than 1.5μ, if this condition is met then we suppress all scale responses that lay beyond this gradient. When the new information introduced by a larger neighbourhood is mostly unresponsive, we see a sharp increase in scale response. The response scale gradient suppression process stops the system assigning a large neighbourhood around a moving entity that is moving through an empty space.

After gradient suppression is applied we are left with a response volume that stores the response at each point across various scales. We perform space-scale Non-Maximum Suppression using a $3 \times 3 \times S$ neighbourhood and locate points of maxima to determine the location and scale of interest points, where S is equal to the number of scale sizes used within the scale selection process.

In addition to suppressing responses in the aforementioned manner we also generate a background mask by computing the absolute inter-frame difference of pixel intensities and applying a fixed threshold of 0.04. This is a necessary step as compression artefacts often result in anomalous optical flow vectors, that when propagated through our proposed method yield a high response where no noticeable movement occurs.

2.6 Feature Extraction

Once we have determined the location and scale of points of interest, we then extract the velocity and acceleration data from trajectories that were initialized within the $N \times N$ area in space as dictated by the selected scale. In addition to this we also extract the data from the response maps R_c and R that fall within the previously defined region. We generate a histogram for each type of data by placing the values into 20 uniformly spaced bins before applying L_2 normalization. Figure 1 shows the locally responsive regions that are detected in scenes taken from the Hockey violence and Violent Flows datasets.

Figure 1. These are local response regions resulting from our process when applied to a sample from the Violent Flows dataset, and the Hockey violence dataset.

3 Experiments

3.1 Datasets

We evaluate our proposed approach using four datasets, two of which were obtained from local police organisations within the United Kingdom; these will be referred to as the CF-violence and NN-violence, named after their respective postcode area. The remaining two datasets are the widely used Hockey violence dataset [14] and Violent Flows dataset [8].

The hockey violence dataset consists of 1000, 50 frame length videos recorded at 25 frames per second at a resolution of 720×576. This dataset displays one-on-one violence between two ice hockey players and footage of standard play. Due to the nature of the sport, the violent acts depicted contain only upper body violence such as punching and pushing. Figure 2 shows examples of the type of footage seen in the hockey violence dataset.

Figure 2. Example frames taken from the Hockey Violence dataset.

The Violent Flows dataset consists of 246 real-life videos

depicting violent and non-violence crowded scenes, and was downloaded from YouTube. Given that all of these videos originate from different sources, they all show different characteristics based not just on crowd behaviour, but also in camera quality and the compression methods applied before being uploaded online. Figure 3 shows samples from the Violent Flows dataset.

Figure 3. Example frames taken from the Violent Flows crowd violence dataset.

The NN-violence dataset contains footage extracted from 18 recorded incidents. The data depicts not just violence but also the scenes before and after a violent incident takes place, therefore providing footage of normal and aggressive behaviour that prefaced the violence. For the purposes of this study, aggressive behaviour will be treated as violent as it can be considered an outlier behaviour that is equally valuable to a CCTV observer as violence. The NN-violence dataset depicts scenes at different times of the day and night resulting in a wide variety of normal pedestrian formation. The footage is recorded at 30 frames per second across a range of cameras with varying quality.

The CF-violence dataset focuses entirely on the Night Time Economy (NTE) and is recorded at a much lower frequency of six frames per second. The CF-violence dataset originates from a more populated city with an arguably bigger NTE which is reflected in the footage, scenes in the CF-violence dataset generally depict denser crowds than those presented in the NN-violence dataset.

Method	Classifier	ROC	Accuracy
Proposed	Random Forest	0.91	84.6 ± 4.1
	Linear SVM	0.93	87.4 ± 1.9
Fast Fight	Random Forest	0.90	82.4 ± 0.6
ViF	Linear SVM	0.88	81.6 ± 0.2
OViF	Linear SVM	0.90	84.2 ± 3.3
MoSIFT	Linear SVM	0.99	96.7 ± 0.7

Table 1. Hockey Violence classification results.

3.2 Experimental Setup

We evaluate the ability to detect violent behaviour using a binary classification approach to separate feature vectors into one of two classes, violent and non-violent. Results are presented as overall classification accuracy and receiver operating characteristic scores. We used five-fold cross validation when evaluating the NN-violence and CF-violence datasets. We perform this process five times and return the average score. Our methods replicate the published evaluation methods used with the Hockey Violence [14] and Violent Flows [8] datasets.

Tests using the proposed approach are conducted using two classifiers for comparison, these are Linear SVM and a Random Forest classifier initialized with 50 trees.

At each frame of a video our proposed approach performs feature extraction using trajectories of length τ. We then apply K-means clustering using a training set of features to form a visual vocabulary. During experimentation we chose the value of 1000 for the vocabulary size for each of the datasets; this value can be further optimized. Each frame in a video is then represented by an L_2 normalized histogram of word occurrences based on the features that occurred in the past W_t frames. A small value for W_t may not capture the co-occurrence of features in time and subsequently miss important relationships between features. In the case of the CF-violence and NN-violence datasets, the value of W_t is set to twice their respective frame-rate. One key variable in our proposed approach is the number of frames we advect our particle, otherwise referred to as trajectory length τ throughout this paper. Given that violent actions are usually characterized by short-term motion then a small value for τ should be adequate to capture the properties of violent behaviour. We found that a value of 8 performed well across all datasets and allowed for real-time feature extraction. In order to reduce computation time we resize all videos so that their spatial dimensions are 160×120. The set of scales used in our multi-scale scheme are $\{12, 24, 36, 48, 60, 72, 84, 96, 108, 120\}$. The weights for each response map (w_1, w_2) were each assigned a value of $\frac{1}{3}$ so that each response type is treated equally and that the final response R is bounded between $[0, 1]$. A future development would be to computationally tune these weights based on the data being analysed. We also use a response threshold of 0.3 to suppress weaker responses.

Our method was implemented with C++ and CUDA and using the previously defined parameters operates at ≈ 44 frames per second when operating on a Nvidia 760 GTX GPU, and i7-4790 CPU at 3.60Ghz.

Method	Classifier	ROC	Accuracy
Proposed	Random Forest	0.87	81.2 ± 3.1
	Linear SVM	0.89	81.7 ± 3.7
Fast Fight	Random Forest	0.75	69.4 ± 5.0
ViF	Linear SVM	0.88	81.2 ± 1.79
OViF	Linear SVM	0.81	76.8 ± 3.9
MoSIFT	Linear SVM	0.88	83.4 ± 8.0

Table 2. Violent Flows classification results.

Method	Classifier	ROC	Accuracy
Proposed	Random Forest	0.91	78.15 ± 3.6
	Linear SVM	0.91	79.4 ± 2.7
Fast Fight	Random Forest	0.89	85.4 ± 5.0
ViF	Linear SVM	0.80	64.6 ± 6.4
OViF	Linear SVM	0.76	59.2 ± 8.6

Table 3. CF-Violence classification results.

Method	Classifier	ROC	Accuracy
Proposed	Random Forest	0.86	64.9 ± 2.7
	Linear SVM	0.82	76.2 ± 3.6
Fast Fight	Random Forest	0.63	60.1 ± 1.1
ViF	Linear SVM	0.70	64.9 ± 3.2
OViF	Linear SVM	0.66	61.2 ± 2.3

Table 4. NN-Violence classification results.

3.3 Results

We compare our results with those obtained using the Violent Flows, Oriented Violent Flows [8], Fast Fight detection [7] and MoSIFT [14] methods of violence detection. Experimentation has shown that our proposed approach, in general, offers good all round classification performance when trying to determine whether or not a scene displays violent behaviour. Table 1 indicates that our approach achieves comparable performance against existing methods when classifying one-on-one violence as shown in the Hockey violence dataset. The application of our proposed method on the Violent Flows dataset, like before, achieves comparable results with existing methods. The ViF and OViF descriptors are extracted globally from the Violent Flows dataset as described in their respective papers, and therefore encode global spatial structure which can result in poor performance when footage is subject to rotation and translation. The application of ViF and OViF on the NN-violence and CF-violence datasets utilize the same approach used to test the Hockey violence dataset, as explained in their respective papers. To give an overview, the authors apply their feature description to a set of interest regions identified using the Space Time Interest Point detector [10]; a Bag of Words model is generated and classification is performed based on feature occurrence within a spatio-temporal volume. Local feature description is preferable over global approaches when analysing surveillance footage as the unfolding violent event is not guaranteed to be centrally viewed by the camera. We demonstrate

Method	Point Detector	Classifier	ROC	Accuracy
ViF	Proposed	SVM	0.89	76.81 ± 4.6
OViF	Proposed	SVM	0.82	65.7 ± 9.0
ViF	STIP	SVM	0.80	64.6 ± 6.4
OViF	STIP	SVM	0.76	59.2 ± 8.6

Table 5. The results obtained when applying ViF and OViF descriptors to regions identified using our proposed solution and STIP.

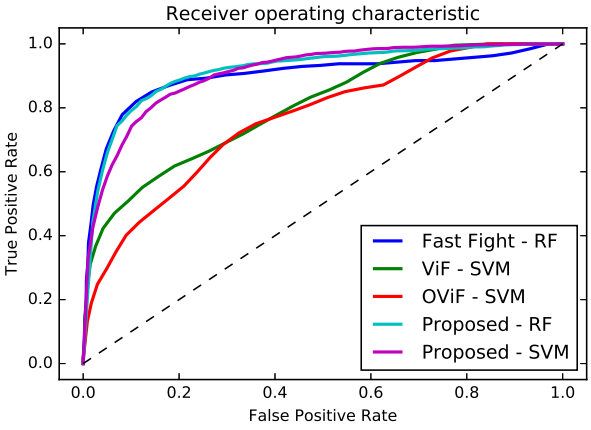

Figure 4. Receiver operating characteristic curve for each method tested on the CF-violence dataset

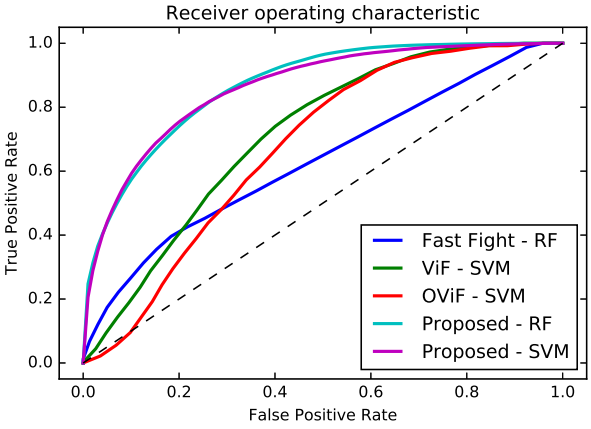

Figure 5. Receiver operating characteristic curve for each method tested on the NN-violence dataset

that our proposed approach outperforms all other tested methods when applied on the city centre surveillance datasets, CF-violence and NN-violence, this is shown in Figure 4 and Figure 5.

We performed a second experiment using the CF-violence dataset in which we described the interest regions found using our proposed approach using the ViF and OViF descriptors. An increase in the ROC score is observed when compared to the same description of regions identified by STIP (Table 5); this suggests that in comparison to STIP, our method of region extraction identifies more relevant parts of a scene when analysing violent behaviour.

4 Conclusion

In this paper we propose three measures of motion trajectory that when combined achieve a maximal response at regions of potentially violent behaviour. A local relative response scheme is applied to achieve scale invariance in order to allow for successful scene description regardless of the distance

between the actors and the camera. We have demonstrated that our approach achieves good performance across a wide variety of violence and offers comparable results on openly available Hockey Violence and Violent Flows datasets. Furthermore, we introduced two datasets and evaluated the ability of multiple methods at detecting violent behaviour captured by real-world surveillance systems.

References

[1] A. Datta, M. Shah, and N. Da Vitoria Lobo. Person-on-person violence detection in video data. In *IEEE International Conference on Pattern Recognition*, volume 1, pages 433–438, 2002.

[2] O. Deniz, I. Serrano, G. Bueno, and T-K. Kim. Fast Violence Detection in Video. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 478–485, 2014.

[3] C. Florence, J. Shepherd, I. Brennan, and T. Simon. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. *BMJ (Clinical research ed.)*, 342:d3313, 2011.

[4] Y. Gao, H. Liu, Xi. Sun, C. Wang, and Y. Liu. Violence detection using Oriented VIolent Flows. *Image and Vision Computing*, 48-49(2015):37–41, 2016.

[5] G. Gerrard and R. Thompson. Two million cameras in the UK. *CCTV Image Magazine*, (42):10–12, 2008.

[6] M. Gill and A. Spriggs. Assessing the impact of CCTV. *Home Office Research, Development and Statistics Directorate*, (February):160, 2005.

[7] I. S. Gracia, O. D. Suarez, G. B. García, and T-K. Kim. Fast Fight Detection. *PLOS ONE*, 10:1–19, 2015.

[8] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[9] D. Helbing, A. Johansson, and H. Z. Al-Abideen. Dynamics of crowd disasters: An empirical study. *Phys. Rev. E*, 75:046109, Apr 2007.

[10] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8. IEEE, 2008.

[11] A. Mecocci and F. Micheli. Real-time recognition of violent acts in monocular colour video sequences. *Signal Processing Applications for Public Security and Forensics, 2007. SAFE '07. IEEE Workshop on*, pages 1–4, 2007.

[12] R. Mehran, A. Oyama, and M. Shah. Abnormal Crowd Behaviour Detection using Social Force Model. *IEEE Conference on Computer Vision and Pattern Recognition*, (1):935–942, 2009.

[13] Y. Nam and S. Hong. Real-time abnormal situation detection based on particle advection in crowded scenes. *Journal of Real-Time Image Processing*, 10(4):771–784, 2015.

[14] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence Detection in Video Using Computer Vision Techniques. *International conference on Computer analysis of images and patterns*, pages 332–339, 2011.

[15] C. Norris and G. Armstrong. *The Maximum Surveillance Society: The Rise of CCTV*. Oxford, 1999.

[16] Great Britain. House of Commons. Home Affairs Committee. A Surveillance Society? London: The Stationery Office Limited, 2008.

[17] P.C. Ribeiro, R. Audigier, and Q. Cuong. RIMOC , a feature to discriminate unstructured motions : Application to violence detection for video-surveillance. *Computer Vision and Image Understanding*, 144:121–143, 2016.

[18] V. Sivarajasingam, J. P. Shepherd, and K Matthews. Effect of urban closed circuit television on assault injury and violence detection. *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention*, 9:312–316, 2003.

[19] G. Van Voorthuijsen, H. Van Hoof, M. Klima, K. Roubik, M. Bernas, and P. Pata. CCTV effectiveness study. *Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology*, pages 5–8, 2005.

[20] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao. Violent video detection based on MoSIFT feature and sparse coding. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3538–3542, 2014.

[21] W. Yu and A. Johansson. Modeling crowd turbulence by many-particle simulations. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(4):1–11, 2007.

[22] X. Zhu, J. Liu, J. Wang, W. Fu, and H. Lu. Weighted interaction force estimation for abnormality detection in crowd scenes. *Computer Vision – ACCV 2012: 11th Asian Conference on Computer Vision*, pages 507–518, 2013.