Violence Detection in Crowded Scenes using Substantial Derivative

Sadegh Mohammadi Hamed Kiani Alessandro Perina Vittorio Murino

Pattern Analysis and Computer Vision Department (PAVIS)
Istituto Italiano di Tecnologia
Genova, Italy

Abstract

This paper presents a novel video descriptor based on substantial derivative, an important concept in fluid mechanics, that captures the rate of change of a fluid property as it travels through a velocity field. Unlike standard approaches that only use temporal motion information, our descriptor exploits the spatio-temporal characteristic of substantial derivative. In particular, the spatial and temporal motion patterns are captured by respectively the convective and local accelerations. After estimating the convective and local field from the optic flow, we followed the standard bag-of-word procedure for each motion pattern separately, and we concatenated the two resulting histograms to form the final descriptor. We extensively evaluated the effectiveness of the proposed method on five benchmarks, including three standard datasets (Violence in Movies, Violence In Crowd, and BEHAVE), and two new video-surveliance sequences downloaded from Youtube. Our experiments show how the proposed approach sets the new state-of-the-art on all benchmarks and how the structural information captured by convective acceleration is essential to detect violent episodes in crowded scenarios.

1. Introduction

With the rapid increasing of surveillance cameras, the analysis of human behavior has attracted a lot of attention in the computer vision community and automatic systems to deal with the scarcity of trained personnel and the natural limitation of human attention capabilities [11].

The biggest challenge of abnormality detection lies in the definition of abnormality as it is strongly context dependent and defined as “outlier” of a normal situation [17]. In the context of video surveillance, examples of abnormalities may be panic or violence. In some contexts, however, people running or walking in some areas or direction of a scene may considered abnormal events.

The most popular approach to detect abnormal behaviors from surveillance camera footages, is to model normality which indeed is a better defined concept. For example, data driven approaches [13, 14, 15] exploit the abundance normal footage to automatically learn a model of ordinary behavior. In alternative approaches the normal behavior is codified by means of a sociological model, being the most famous example the social force model [9].

A diametrically different approach is to focus on abnormalities. Clearly, they do not represent a compact or well defined concept. Thus, one has to restrict to a particular case like panic [8], violence [10] or drunkiness [16]. This restriction may lead to a better performance simply because of the clear definition of the task. It is also important to note that in this case *i)* despite of scarcity, abnormal footage may be available and discriminative methods like support vector machines can be applies, and *ii)* one can focus on the characteristics of the abnormality and design a feature representation which exhibits the discriminative patterns (mainly in terms of motion and appearance) of normal and abnormal activities. This paper takes the latter approach and focuses on the automatic classification of violence episodes in crowded scenarios.

Violence detection in video sequences is not a novel problem. Despite recent improvements, effective solutions for real-worlds situations are still unavailable. The first paper appeared on this topic is [5] which focuses on two persons fight episodes and uses motion trajectory information of person’s limbs for classification. Besides, only focusing on person-on-person interactions, this method requires the segmentation of the silhouette, consequently, it is not easily exploitable in crowded scenarios.

More robust methods [6, 18, 7] only focus on visual cues and they are all based on the “bag of words” paradigm. The differences between [6, 18, 7] lie in the sampling strategy, the feature descriptor or the classifier used. For example [6] used STIP detector and descriptor and linear support vector machines. The approach proposed in [18] employed STIP detection and HOG/MoSIFT descriptor along with the histogram intersection kernel while [7] used random sampling

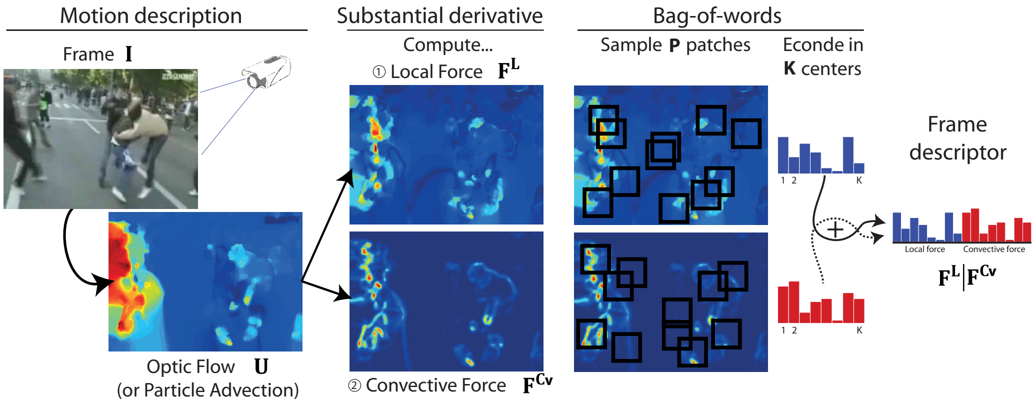

Figure 1. Overview of the proposed framework.

and optical flow magnitude. The success of each of these method depends on the frame quality and the density of the people involved in the violent act. Simplest descriptors based on optic flow and dense sampling, however, performed fairly well in all scenarios and offered quasi real-time performance as shown in [7].

From this brief literature review emerges that the bag-of-words paradigm preforms well being especially robust to crowded scenarios. Moreover, the motion descriptors tailored for action recognition often fail for the task of abnormality detection in crowd due to the unpredictable and sudden changes in crowd motions, which are specific characteristics of violences or riots. Our intuition is that a crowd descriptor should capture the *changes in motion* more than the motion itself which can be captured by higher order derivatives.

This paper proposes a novel computational framework for detecting abnormal events in video sequences. We exploit the use of substantial derivative, an important concept in fluid mechanics which encompasses *spatial* and *temporal* information of motion changes in a single framework [2]. In a nutshell, the substantial derivative equation captures two important properties: *i) local acceleration* which is velocity changes with respect to time at a given point and occurs when the flow is unsteady and *ii) Convective acceleration* which is associated with spatial gradients of velocity in the flow field. Convective acceleration occurs when the flow is non-uniform and its velocity changes along is trajectory. Particularly, it is useful to capture useful information in the crowd scenario where size of parties participating in the violent is not uniform and structure of motion varies drastically.
Our framework is summarized by Figure 1. First, we extract a motion description by the means of dense optic flow (particle advection [1] can also be used). Second, following the substantial derivative equation, we compute *local force* and *convective force* between each consecutive pairs of frames. Then, we followed the standard bag-of-words paradigm for each force separately, sampling P patches and encoding them in K centers. Finally, we concatenate the histograms to form the final descriptor, which we defined as *total force*.

The rest of the paper is organized as following. Section 2 covers the main concepts of fluid dynamics and the substantial derivative equation, also discuses the parallel between fluids and crowds. In Section 3 we show how the substantial derivative equation can be employed to extract motion primitives and finally in Section 4 we present an exhaustive experimental section.

2. The Substantial Derivative Model

In this section, first, we introduce the main concepts behind the substantial derivative in fluid dynamics, then we discuss about its effectiveness to offer discriminating features to distinguish violent behaviors in crowded scenarios.

2.1. The substantial Derivative in Fluid Mechanics

Substantial derivative is an important concept in fluid mechanics which describes the change of fluid elements by physical properties such as temperature, density, and velocity components of flowing fluid along its trajectory $(\mathbf{p}, t)$ [2]. In particular, given the velocity components of a certain flowing particle in the x- and y-direction in the time t, its velocity flow evolution along its trajectory $\mathbf{U} = U(\mathbf{p}, t) = U((x, y), t)$ can be described as:

$$\begin{aligned} \frac{D_{\mathbf{U}}}{D_t} &= \frac{\partial_{\mathbf{U}}}{\partial t} + u\frac{\partial_{\mathbf{U}}}{\partial x} + v\frac{\partial_{\mathbf{U}}}{\partial y} \\ &= \frac{\partial_{\mathbf{U}}}{\partial t} + (\mathbf{U} \cdot \bigtriangledown)\mathbf{U} \end{aligned} \quad (1)$$

Where $\frac{D_{\mathbf{U}}}{Dt}$ is the substantial derivative and indicates the *total acceleration* of the flowing particle moving along its trajectory. The term $\frac{\partial_{\mathbf{U}}}{\partial_t}$ computes the *local acceleration*.

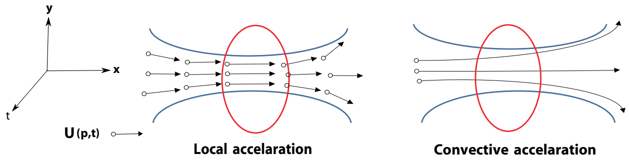

Figure 2. An example of local and convective accelerations. The local acceleration measure instantaneous rate of change of each fluid particle, while convective acceleration measures the rate of change of the particle moving along its trajectory. Red region indicates the particles are accelerated as it converge due to the structural change of the environment.

$\frac{\partial \mathrm{U}}{\partial_x}$ and $\frac{\partial \mathrm{U}}{\partial_y}$ are, respectively, the partial derivative of velocity field $\mathbf{U}$ to the x and y directions. $u = \frac{\partial_x}{\partial_t}$ and $v = \frac{\partial_y}{\partial_t}$ are the velocity components of the particle in the x- and y-direction respect to the time t. Finally $(\mathbf{U} \cdot \nabla)\mathbf{U}$ computes the *convective acceleration* where $\nabla \equiv \frac{\partial}{\partial x} + \frac{\partial}{\partial y}$ is the divergence operator.

The local acceleration captures the change rate of velocity of a certain particle respect to time and vanishes if its flow is steady. The convective acceleration, on the other hand, captures the change of velocity flow in the spatial space and, therefore, it increases when particles move through the region of spatially varying velocity. In this case, one can say that the local acceleration characterizes the particle velocity field in the temporal domain, while the convective acceleration represents the velocity change due to the spatial variation of the flow particle along its trajectory (see Figure 2 for an example). We incorporate both convective and local accelerations to model the pedestrians motions dynamics in crowd scenes.

2.2. Modeling Pedestrian's Motion Dynamics

In the following, we describe how the substantial derivative can be applied to model pedestrian motion dynamics in a crowd. Suppose that M pedestrians with mass of $m_i, \quad i = 1, .., M$ and corresponding velocities v_i are involved in a crowd. The total force that govern the motion of each pedestrian is $F_i^T = m_i \cdot a_i^T$, where $a_i^T = a_i^L + a_i^{Cv}$ is the total acceleration, a_i^L and a_i^{Cv} are respectively the local and convective accelerations, and m_i is the mass of the pedestrian i. Therefore, the total force of pedestrian/particle i can be expressed as:

$$F_i^T = m_i \cdot a_i^T = F_i^L + F_i^{Cv} \tag{2}$$

where $F_i^L = m_i \cdot a_i^L$ and $F_i^{Cv} = m_i \cdot a_i^{Cv}$ are indicated as the local and convective forces, respectively, caused by the local and convective accelerations.

In abnormal scenarios such as violence, in particular, an individual shows intentional aggressive behaviors against another one in with a sudden change of his/her velocity field in the temporal domain (time). This crowd motion pattern can be characterized by the local force F^L . Moreover, the motions of people involved in an abnormal crowd, e.g. violence, are convened by the crowd dynamics and mainly are unpredicted and sudden. These motion changes show the spatial gradients of velocity fields of people within an abnormal crowd situation which can be represented by the convective force F^{Cv}. By integrating the local and convective forces F^L and F^{Cv} into the total force F^T, we are able to simultaneously capture the spatial structure and temporal changes of motion fields within video sequences.

3. Estimation of local and convective forces from videos

In this section, we detail the process of estimation of local and convective forces from a video.
As a first step, we computed the optic flow of the video sequences using the algorithm presented in [12] (any other method can be employed). For each frame of the video $\{I^t\}_{t=1}^N$, the optic flow $\{\mathbf{U}^t\}_{t=1}^{N-1}$ represents an estimate of the velocity components of each pixel in the x and y direction, e.g., $\mathbf{U}^t(x, y) = (v_x^t, v_y^t)$.

According to (1), the *local acceleration* a^L is the derivative of the velocity with respect to the time. By considering a unit time change (per frame), the local acceleration in the x and y directions of two consecutive optical flows can be computed by

$$a_x^t = v_x^t - v_x^{t-1} \quad \text{and} \quad a_y^t = v_y^t - v_y^{t-1} \tag{3}$$

Given the two components a^x and a^y, we extract the magnitude of a^L as $a^L = \sqrt{(a_x)^2 + (a_y)^2}$, this is shown in Figure 3.

The *convective acceleration* a^{Cv} captures the spatial evolution of a particle along its trajectory. This requires to track each particle (individuals in our case) both in the spatial and temporal domain. Tracking individuals, however, is a very challenging task especially in crowded scenarios and likewise previous work, we resort to particle advection [14]. Following the standard procedure, we placed an homogeneous grid of particles over the video frames and we "advected" them according to the average optic flow over a fixed window of time and as well as space. This is done by a weighted average using a gaussian kernel. Using the described process, each particle moves with the average velocity of their neighborhood, resembling the collective velocity of a group of people in the crowd.

Given the averaged velocity components that move each particle, e.g., $\{\bar{v}_x^t, \bar{v}_y^t\}_{t=1}^{N-1}$, we compute their spatial derivatives in the x and y directions the convective acceleration

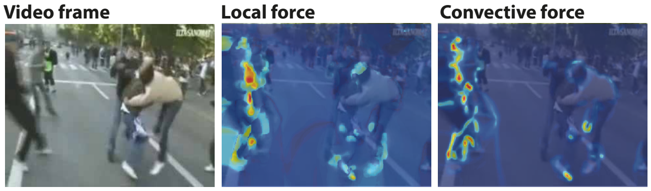

Figure 3. Examples of computed local and convective force fields for a video sequence. The image on the left is the video frame, the heat map in the middle is computed local force and on the left is convective force underplayed over original frame. Red pixel corresponds to the higher force values.

components

$$\bar{a}_x = (\frac{\partial \bar{v_x}}{\partial x} + \frac{\partial \bar{v_y}}{\partial y}) \cdot \bar{v}_x \quad \text{and} \quad \bar{a}_y = (\frac{\partial \bar{v_x}}{\partial x} + \frac{\partial \bar{v_y}}{\partial y}) \cdot \bar{v}_y \quad (4)$$

Then, the magnitude of the convective acceleration is computed by $a^{Cv} = \sqrt{(\bar{a}_x)^2 + (\bar{a}_y)^2}$. The convective acceleration for a particular frame is shown in Figure 3; for a better visualization we computed it using dense optic flow (e.g., no particle advection) using the same procedure of local acceleration. Finally, following the prior work [14], if we assume that all individuals in a crowd have a unit mass of $m_i = 1$, in this case local and convective forces are respectively equal to the local and convective accelerations, $F^L = a^L$ and $F^{Cv} = a^{Cv}$.

Given convective and local forces computed for each video, we applied the standard bag-of-words method separately for local and convective forces. For each video we randomly sampled P patches of size $5 \times 5 \times 5$ and we learned a visual dictionary of size $K = 500$ cluster centers using K-means[1]. In the bag-of-word assumption each video is encoded by a bag; to compute such bags we assigned each of the P patch to the closest codebook and we pooled together all the patches to generate an histogram over the K visual words. The final descriptor is simply computed by concatenating the histograms of local and convective forces. With a little abuse of notation, in the experiments we will refer to these histogram-descriptors computed from local and convective force with F^L and F^{Cv} and to the final descriptor which we refer to as $F^L|F^{Cv}$

4. Experimental Setup

We evaluated our approach on three standard benchmarks namely *Violence in Movies* [18], *Violence in Crowds* [7] and *BEHAVE* [3], and two new sequences downloaded from `www.youtube.com` that we named *Panic* and *Riot in Prison*. Figure 4 shows few frames for each dataset; as visible each dataset reflects different conditions in terms of number of people involved, camera motion and view-point.

Violence In Movie consists of 200 short videos, including 100 person-on-person fight, collected from action movies and 100 non-fight scenarios obtained from action recognition datasets. This is a challenging dataset, because of the variety of scenes and imaging conditions.
Violence in Crowds [7] is a new benchmark specifically assembled for violence classification in crowded scenarios like stadiums, rallies or demonstrations. In total, it contains 246 video sequences: 123 normal and 123 violent.
BEHAVE dataset [3] was collected from a surveillance camera under experimental conditions. The dataset contains different group activities, including approaching, walking together, meeting, splitting, ignoring, chasing, following, running together, and fighting with a number of participants varying from 2 to 7 pedestrians. Similarly to [4], we divided the videos into two classes, considering fighting as abnormal behaviors and the rest as normal.
Riot In Prison is a video sequence recorded with a surveillance camera inside a prison. After several normal frames, multiple person-on-person fights occur, with the number of participants increasing gradually. Finally, security guards intervene and the sequence ends with a normal situation. This sequence is 3728 frames long, being 1160 the violent frames.
Panic . Likewise the previous one, we downloaded this sequence form youtube. It consists with 2207 frames including 1962 frames of normal and 245 abnormal. The video was recorded with moving camera in outdoor with day light illumination. We considered this sequence to evaluate the robustness of the proposed descriptor to a different abnormality e.g., panic. It is important to note the different nature of the datasets we considered. Violence in movies and violence in crowds are standard benchmarks for video-level violence classification, where violent data is available at training time. Following the standard bag-of-words paradigm, we described each video with a bag and we employed SVM with Histogram intersection kernel [19] as a classifier. Results are presented in terms of classification accuracy.
For the remaining three sequences the goal is temporal detection. We divided each sequence in temporally overlapping clips of 15-frames length with 5 frames of overlapping, we described each clip with a bag and we tried to detect violence at clip-level. In this case abnormal data is not available and we resorted to a standard data driven approach: firstly we learned a latent Dirichlet allocation model to encode the normal behavior, then we evaluated and thresholded the likelihood of each test clip to decide about the normality of a clip. Results are presented in terms of area under the ROC curve (AUC). In order to

[1] To employ k-means, we rasterized each patch in a vector of size 125 and we used euclidean distance

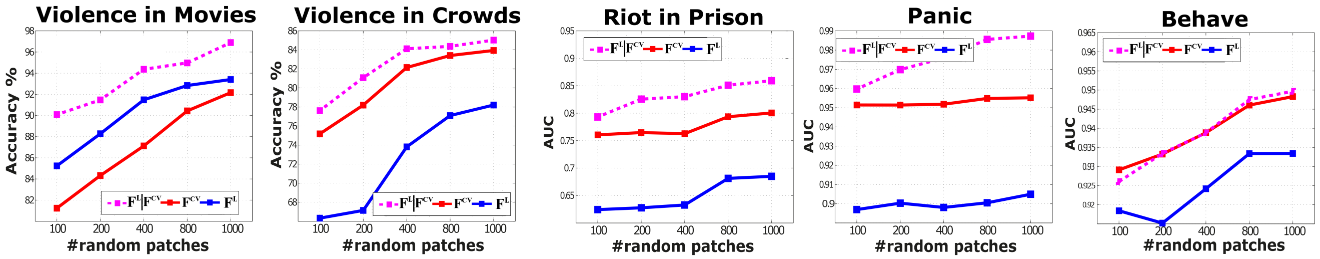

Figure 5. Comparison of average accuracy of proposed forces on VIM, and VIC using SVM with 5-fold cross validation, and AUCs of proposed forces on Riot In Prison, Panic, and Behave sequences varying the number of sampled patches.

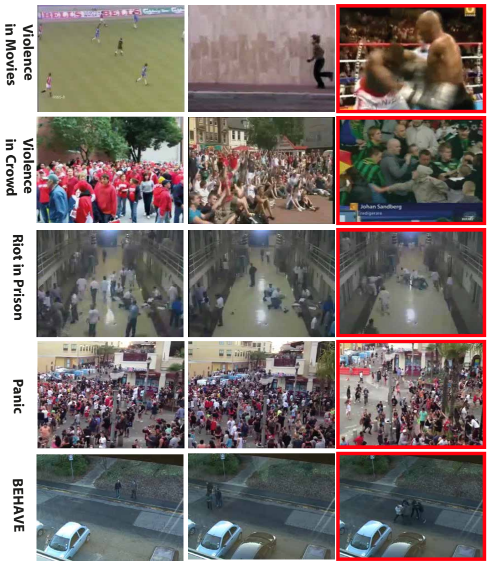

Figure 4. Two normal frame and one abnormal such as fight or panic frame (right-most column) for each of the datasets considered in our study.

compensate the effect of the random sampling, we run each test with 10 repetitions and we reported the mean accuracy - AUC along with the 95% confidence interval.

In the first experiment, we examined the effect of the number of patches P used to compute the bags. We randomly sampled $P \in \{100, 200, 400, 800, 1000\}$. Figure 5 shows how the overall performance varies with P for local F^L, convective F^{Cv} and the total force obtained by concatenating local and convective forces.

As expected, for all the datasets the performance improved increasing P and considering combination of the forces. Interestingly one can also observe that in crowded scenarios convective force outperformed significantly the local force; this demonstrates the importance of structural information to uncover underlying dynamics of crowd motion pattern, which results in, more discriminating features.

In the second experiment, we compared our descriptors with the state-of-the-art. As baselines we considered the bag-of-word representation computed from the optic flow (e.g., violent flow [7]), jerk[2] and interaction force [14]. For all our baselines, we used the very same procedure as F^L and F^{Cv}, setting $P = 1000$ and $K = 500$. Finally, as additional comparison we reported results from other papers focusing on action recognition descriptors in violence in movies and violence in crowds dataset.

Tables 1 - 4 report average results and the confidence intervals. As immediately visible our features performed well on each dataset and the total force, set the state-of-the art on each dataset.

Overall, we observed that in the densely crowded situations structural information (F^{Cv}) had a significant effect on the performance of the classifier while in less crowded scenes, temporal (F^L) and structural information preformed almost equally well. Finally, in person-on-person fights, temporal information outperformed structural information, nevertheless their combination significantly improved the overall performance of the classifier.

It is also worth to note that on BEHAVE dataset our method only matched the result of the energy potential [4]. However, as major drawback [4] employs a support vector machine and it requires abnormal data at training time.

Table 1. Average accuracy and 95% confidence interval for the *Violence in Movies* dataset using 5-fold cross-validation. † results taken from [18].

Method	*Accuracy*
STIP (HOF)† [18]	50.5%
MoSIFT† [18]	89.5%
Optic flow (ViF)	91.31 ± 1.06%
Interaction Force [14]	95.51± 0.79 %
Jerk [5]	95.02± 0.56%
Local force - F^L	93.4 ± 1.24%
Convective force - F^{Cv}	92.16± 1.13 %
$F^L\|F^{Cv}$	**96.89± 0.21%**

[2] Jerk or Jolt is the temporal derivative of acceleration. It was the base feature considered in [5]

Table 2. Average accuracy and 95% confidence interval for the *Violence in Crowds* dataset using 5-fold cross-validation. † results taken from [7].

Method	*Accuracy*
HOT† [17]	82.30%
LTP† [7]	71.53 ± 0.15%
Dense Trajectories [20]	79.38± 0.14 %
Optic Flow (ViF)† [7]	81.30± 0.18 %
Interaction Force [14]	74.5 ± 0.65 %
Jerk [5]	74.18± 0.85%
Local force - F^L	78.14± 0.92%
Convective force - F^{Cv}	84.03± 1.34 %
$F^L \vert F^{Cv}$	**85.43± 0.21%**

Table 3. Average AUCs and 95% confidence interval on Riot In Prison and Panic sequences.

Method	*Riot In Prison AUC*	*Panic AUC*
Optic Flow (ViF) [7]	0.76 ± 0.052	0.89 ± 0.0136
Interaction Force [14]	0.66 ± 0.024	0.89 ± 0.0040
Jerk [5]	0.65 ± 0.036	0.90 ± 0.0095
Local force -F^L	0.68 ± 0.027	0.90 ± 0.0079
Convec. force - F^{Cv}	0.79 ± 0.014	0.95 ± 0.0023
$F^L \vert F^{Cv}$	**0.85 ± 0.077**	**0.98 ± 0.0055**

Table 4. Comparison of average AUCs on Behave dataset, with 95% confidence interval. † results taken from [4].

Method	*Classifier*	*AUC*
Energy Potential† [4]	SVM	**0.94**
Interaction Force † [14]	SVM	0.88
Optic Flow (ViF)† [7]	SVM	0.81
Interaction Force [14]	LDA	0.925 ± 0.008
Optic Flow (ViF)	LDA	0.901 ± 0.032
Local force - F^L	LDA	0.933± 0.073
Convective force - F^{Cv}	LDA	0.946 ± 0.032
$F^L \vert F^{Cv}$	LDA	**0.948 ± 0.054**

5. Conclusions

We introduce a novel computational framework to classify violence behaviors in various scenarios. In particular, we addressed the ability of the method to capture the dynamics of pedestrians based on spatial-temporal characteristics of substantial derivative. The results of our method, indicated that the importance of spatial information to reveal complex pedestrian's dynamics in crowded scenarios. We demonstrated that the combination of the spatial and temporal motion patterns mostly have a significant effect on the performance of the classifiers. Finally, our descriptor shows its effectiveness not only in various violent situations, but also considering the panic situation as an abnormal situation.

References

[1] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, pages 1–6. IEEE, 2007.

[2] G. K. Batchelor. *An introduction to fluid dynamics*. Cambridge university press, 2000.

[3] S. Blunsden and R. Fisher. The behave video dataset: ground truthed video for multi-person behavior classification.

[4] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In *CVPR*, pages 3161–3167. IEEE, 2011.

[5] A. Datta, M. Shah, and N. da Vitoria Lobo. Person-on-person violence detection in video data. In *PR*, volume 1, pages 433–438. IEEE, 2002.

[6] F. D. M. de Souza, G. C. Chávez, E. do Valle, D. A. Araujo, et al. Violence detection in video using spatio-temporal features. In *SIBGRAPI*, pages 224–230. IEEE, 2010.

[7] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPRW*, pages 1–6. IEEE, 2012.

[8] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490, Sept. 2000.

[9] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[10] W. Jager, R. Popping, and H. van de Sande. Clustering and fighting in two-party crowds: Simulating the approach-avoidance conflict. *J. Artificial Societies and Social Simulation*, 4(3), 2001.

[11] H. U. Keval. *Effective design, configuration, and use of digital CCTV*. PhD thesis, University College London, 2009.

[12] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Citeseer, 2009.

[13] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, pages 1975–1981. IEEE, 2010.

[14] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942. IEEE, 2009.

[15] S. Mohammadi, H. Kiani, A. Perina, and V. Murino. A comparison of crowd commotion measures from generative models. In *CVPRw*, 2015.

[16] S. C. Moore, M. Flajlik, P. L. Rosin, and D. Marshall. A particle model of crowd behavior: Exploring the relationship between alcohol, crowd dynamics and violence. *Aggression and Violent Behavior*, 13(6):413 – 422, 2008.

[17] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino. Analyzing tracklets for the detection of abnormal crowd behavior. In *WACV*, pages 148–155. IEEE, 2015.

[18] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar. Violence detection in video using computer vision techniques. pages 332–339, 2011.

[19] M. J. Swain and D. H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.