Angry Crowds: Detecting Violent Events in Videos

Sadegh Mohammadi[1(✉)], Alessandro Perina[1,2], Hamed Kiani[1], and Vittorio Murino[1,3]

[1] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, Genova, Italy
{sadegh.mohammadi,vittorio.murino}@iit.it
[2] Microsoft Corp, WDG Core Data Science, Redmond, Italy
alperina@microsoft.com
[3] Department of Computer Science, University of Verona, Verona, Italy

Abstract. Approaches inspired by Newtonian mechanics have been successfully applied for detecting abnormal behaviors in crowd scenarios, being the most notable example the Social Force Model (SFM). This class of approaches describes the movements and local interactions among individuals in crowds by means of repulsive and attractive forces. Despite their promising performance, recent socio-psychology studies have shown that current SFM-based methods may not be capable of explaining behaviors in complex crowd scenarios. An alternative approach consists in describing the cognitive processes that gives rise to the behavioral patterns observed in crowd using heuristics. Inspired by these studies, we propose a new hybrid framework to detect violent events in crowd videos. More specifically, (i) we define a set of simple behavioral heuristics to describe people behaviors in crowd, and (ii) we implement these heuristics into physical equations, being able to model and classify such behaviors in the videos. The resulting heuristic maps are used to extract video features to distinguish violence from normal events. Our violence detection results set the new state of the art on several standard benchmarks and demonstrate the superiority of our method compared to standard motion descriptors, previous physics-inspired models used for crowd analysis and pre-trained ConvNet for crowd behavior analysis.

Keywords: Violent events · Social force model · Behavioral heuristics

1 Introduction

Video surveillance cameras have become ubiquitous in our cities. However, their usefulness for preventing crimes, is often questioned due to the lack of adequately trained personnel to monitor a large number of videos captured simultaneously,

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46478-7_1) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 3–18, 2016.
DOI: 10.1007/978-3-319-46478-7_1

and to the loss of attention from surveillance operators after a few tens of minutes inspecting such videos [1].

This has attracted great attention from the computer vision community aimed at developing techniques to automatically detect abnormal behaviors in videos which may preserve safety and likely prevent crimes.
Although the proposed methods have achieved significant outcomes, still they are far away to being applied for real world scenarios. In particular, the biggest challenge lies in the definition of abnormality as it is strongly context dependent. In most cases, violence and panic are considered as abnormal behaviors, however, even people running or walking in some areas of a scene may be considered an abnormal event in particular situations. In video surveillance scenarios, the most well-known approach to detect abnormalities is to codify the pedestrians' behaviors by means of sociological models, being the most notable example the social force model (SFM) [2], which was successfully employed for abnormality (mostly panic) detection in crowd scenes [3]. Specifically, SFM is a method for describing local crowd interactions using Newtonian mechanics.

Although many variants of the SFM have been proposed in the social psychology literature [4–6], the central tenet of all such models is the ability to describe different crowd scenarios (e.g., cross walk, panic and evacuation) by calibrating a set of physical forces on empirical observations [7]. Despite the interesting performances of the SFM-based models [2], recent social psychology studies argued that they are too simplified [7,8] to capture complex crowd behaviors, other than being heavily affected by a poor generalization power, meaning that a model calibrated on a set of empirical observations may often fail to deal with a different set of observations[1].

To face these limitations, recent works try to exploit a set of simple, yet effective, *behavioral heuristic* to describe complex individuals' behaviors observed in crowded scenarios, while using physics-based equations to quantify such rules on crowd videos [7–9]. Unlike SFM-based models which aim at describing complex crowd movements by calibrating a set of forces on empirical observations, this class of approaches defines a set of behavioral heuristic which are formulated using concepts such as velocity and acceleration borrowed from Newtonian mechanics [7]. The effectiveness of such heuristics for modeling complex human (re)actions and decision-making have been well noted in psychology literature [10–12] and share the common characteristics to be *fast* and *frugal* [13]. They are fast because of their low computational complexity, and frugal since they benefit from a few pieces of information [13]. Readers may refer to [7,8] for a full treatment of the above methods from psychological and sociological perspectives.

In this work, taking inspiration from such socio-psychological studies above mentioned, we propose to employ cognitive heuristics together with physical equations for detecting violence in video sequences. To the best of our knowledge, this is the first attempt in computer vision that investigates the use of heuristic rules for violence detection in crowd scenarios. More specifically, **(I)** We extended

[1] This is referred to low predictive power in socio-psychology [7].

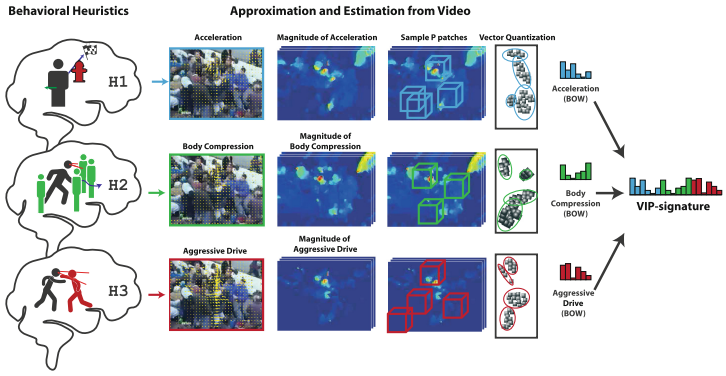

Fig. 1. Overview of the proposed framework, from behavioral heuristic rules to the Vision Information Processing Signature descriptor (VIPS).

the heuristics of the cognitive model proposed in [7,8] to model violent in crowds. **(II)** We formalized the heuristics with mathematical equations and **(III)** we showed how we are able to efficiently approximate and extract them from a video sequence. **(IV)** Finally, we use the estimated heuristic maps to form a video descriptor, called Vision Information Processing Signature (VIPS) which strongly outperform the social force model, ConNet and other state of the art descriptor on the violence classification task.

Figure 1 depicts an overview of our framework. First, we define three behavioral heuristic rules based on social-psychology studies [7,8]. Then, we compute motion information from two successive frames along with particle advection to track particles (to capture as much as possible individual subject motion in crowd scenes). This is followed by computing physics-based feature maps from each behavior heuristic rule. Finally, following the standard bag-of-words paradigm, we sampled P patches and encode them into a number of centers. Then, we concatenate the histograms to form the VIPS descriptor. Eventually, The resulting histograms are fed into a classifier to detect/quantify the violence behaviors.

The rest of the paper is organized as follows. In Sect. 2, we review the state-of-the-art on violence detection using computer vision techniques. Section 3 presents the proposed cognitive models and describe the envisaged heuristic rules. In Sect. 4, we illustrate how to estimate the formulated forces from video sequences. This involves extracting a set of maps from the heuristics, which we will further exploit to define the VIPS descriptor for crowd violence detection. In Sect. 5 we evaluate our approach on several benchmark datasets comparing with prior dominant techniques and descriptors. Finally, Sect. 6 draws a conclusion and presents the future work.

2 Related Works

The first work for detecting violence in videos was proposed in [14]. This approach focused on two- person fight episodes and employed motion trajectory

information of individual limbs for fight classification. It required limbs segmentation and tracking, which are very challenging tasks in presence of occlusion and clutters, specially in crowd situations.

More recent methods [15–19] mainly differ in the used feature descriptor, sampling strategy and the classifier adopted. For example, [15] used Spatial Temporal Interest Point (STIP) detector and descriptor along with linear Support Vector Machines (SVMs). Nievas et al. [20] applied STIPs, Histogram of Oriented Gradients (HOG) and Motion SIFT (MoSIFT) descriptors along with the Histogram Intersection Kernel SVM [20] for violence detection. Other approaches derived local motion patterns from optical flows. For instance, Solmaz et al. [21] analyzed motion flows (derived from optical flows) to identify a particular set of simple crowd behaviors (e.g., bottlenecks and lanes). The statistics of flow-vector magnitudes changing over the time are exploited in [18] to represent motion patterns for the task of violence detection. The social force model [3] and its variations [22–24] represented motion patterns using physics concepts such as attractive and repulsive forces, motion equations and interaction energy. The success of this class of methods, however, is heavily dependent on the video quality and the density of people involved in crowds, and they may not be capable of capturing a wide range of complex crowd behaviors.

3 Formulation of Heuristic Rules

In this section, first, we define a set of heuristic rules inspired from socio-psychological studies [7–9] describing how individuals behave in violence crowd. Then, we explain how to formulate these rules using physics equations and basic visual information extracted from the observed scenes.

Our proposed framework consists of the following heuristic rules:

`H1`: *An individual chooses the direction that allows the most direct path to a destination point, adopting his/her moving regarding the presence of obstacles.*
`H2`: *In crowd situations, the movement of an individual is influenced by his/her physical body contacts with surrounding persons.*
`H3`: *In violent scenes, an individual mainly moves towards his/her opponents to display violent actions.*

The first heuristic rule (`H1`) is inherited from the socio-psychological literature [7] and encompasses individual's internal motivation towards a goal avoiding obstacles or other individuals. The second heuristic rule (`H2`), on the other hand, states that individual movements in a crowd is not only governed by his/her internal motivation but also by the unintentional physical body interactions with his/her surrounding individuals. This is especially true in overcrowded situations where crowd dynamics is unstable and body contacts frequently occur. The third heuristic rule (`H3`) defines behavioral patterns within violent scenes, where there are two or more parties (e.g., police and rioters) fighting and showing violent behaviors to each other.

We formulate the above heuristic rules using visual information of individuals such as their spatial coordinates and velocity flows, following [7–9]. For each individual i, we consider its position (x_i, y_i) in the 2D image plane and its velocity $\boldsymbol{v}_i$. With the scalar $d_{i,j}$, we refer to the distance between i and j, and $\boldsymbol{n}_{ji}$ is a normalized unit vector pointing from the coordinates of j to i. The visual motion information of i with respect to j is captured by the angle between the velocity vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, which we call it ϕ_{ij}. Based on these visual cues, the heuristic rules are formulated as follows.

Heuristic rule `H1`: In normal situations, individual i chooses the most direct path towards a destination with a desired velocity of $\boldsymbol{v}_i^{des}$. It is, however, a norm that individual i changes his/her desired velocity $\boldsymbol{v}_i^{des}$ to $v_i(t)$, due to an unexpected obstacle at time t [2]. This heuristic can be formulated as:

$$\frac{d\boldsymbol{v}_i^{des}}{dt} = \frac{(\boldsymbol{v}_i^{des} - \boldsymbol{v}_i(t))}{\tau} \tag{1}$$

where τ is the amount of time individual i requires to change its desired velocity facing an unexpected obstacles. If velocity is constant over time, $\frac{d\boldsymbol{v}_i^{des}}{dt} = 0$, meaning the individual is approaching his/her target destination without facing any obstacle. Otherwise, the presence of an obstacle implies a change at the individual 's velocity.

Heuristic rule `H2`: The heuristic `H1` is, however, valid in sparse crowd scenarios (e.g., walking in a street) where individuals have enough time and space to keep safe distance from other pedestrians, and change their desired velocity against unexpected obstacles. This is not the case in crowd situations (e.g., riots), where individuals do not have enough time and space to control their movements. Hence, they are subject to unintentional physical body contacts that may strongly affect their movements. Borrowing from [7,25], the **b**ody **c**ontact force imposed on i from j is formulated as:

$$\boldsymbol{F}_{ij}^{bc} = \boldsymbol{n}_{ji} \cdot \mathbf{g}_i(j) \tag{2}$$

where $\mathbf{g}_i(j)$ is a function that returns zero if i and j are not close enough to have body contact and a scalar value inversely proportional to their spatial distance d_{ij}, otherwise.

Heuristic rule `H3`: In violent situations, individual j may exhibit an action (verbally, emotionally or physically) to individual i that triggers i to move towards j for a reaction [26]. This is what heuristic `H3` aims to model. We named this as aggression force $\boldsymbol{F}_{ij}^{agg}$ which is defined as:

$$\boldsymbol{F}_{ij}^{agg} = \boldsymbol{n}_{ji} \cdot \frac{(1 - \frac{\boldsymbol{v_i} \cdot \boldsymbol{v_j}}{\|\boldsymbol{v_j}\| \cdot \|\boldsymbol{v_i}\|})}{2} \cdot \mathbf{f}_i(j) \tag{3}$$

$\mathbf{f}(\cdot)$ returns 1 for each individual j who is in the view field of individual i regardless of their distance, and 0 otherwise. The term $\frac{1}{2}(1 - \frac{\boldsymbol{v_i} \cdot \boldsymbol{v_j}}{\|\boldsymbol{v_j}\| \cdot \|\boldsymbol{v_i}\|})$ is referred to as aggression factor and measures how much the individual i is stimulated

to move towards j based on the angle between the velocity vectors $\boldsymbol{v_j}$ and $\boldsymbol{v_i}$. The value of aggression factor is in the [0 1] interval, 1 when individuals i and j are moving against each other (the angle between vectors $\boldsymbol{v_j}$ and $\boldsymbol{v_i}$ is π), and 0 when individuals i and j are moving towards a same direction (the angle between $\boldsymbol{v_j}$ and $\boldsymbol{v_i}$ is 0). $\boldsymbol{n}_{ji}$ codes the spatial relation of i and j and gives the aggression force a vector form (with direction and magnitude).

4 Estimating Heuristic Rules from Videos

In this section, we quantify each heuristic rule on video sequences. This provides a set of maps (one map for each rule) which will be further used to define our video descriptor for violence detection.

Assume that the goal is to quantify the heuristic rules on a gray-level video $\mathbf{V} = \{I^1, ..., I^T\}$ with T frames of size $h \times w$. Toward this, we need to compute the basic variables in Eqs. 1–3, including each individual's spatial coordinate and velocity. This can be performed by detecting and tracking individuals over the video frames. This, however, is very challenging in crowd videos with severe occlusions and clutter. An alternative, without individual detection and tracking, is particle advection [3], where a grid of particles is placed over each frame and moved according to the video flow field computed from the optical flow (OF) [27]. The velocity vector of each particle i located at (x_i, y_i) over frame t is approximated by averaging OF vectors in its neighborhood using a Gaussian kernel in the spatial and temporal domains, i.e., $\boldsymbol{o}_i = \left\langle \mathbf{OF}(x_i, y_i, t) \right\rangle_{avg}$. More details about particle advection can be found in [3]. From now on, we will use the term particle(s) instead of individual(s), and optical flow instead of velocity.

Estimation of heuristic rule `H1`. The formulation of heuristic rule `H1` estimates the change of a particle's velocity over the time, which is particle's *acceleration*, $\boldsymbol{a_i}$[2]. Borrowing from [3], we estimate Eq. 1 by computing the derivation of OF vectors with respect to the time:

$$\frac{d\boldsymbol{v}_i}{dt} = \frac{\boldsymbol{v}_i^{des} - \boldsymbol{v}_i(t)}{\tau} \simeq \frac{\boldsymbol{o}_i^{t+\Delta t} - \boldsymbol{o}_i^{t}}{\Delta t} \tag{4}$$

where the apex states the time (frame index). If we set $\Delta t = 1$ (two successive frames), then the particle's *acceleration* ($\frac{d\boldsymbol{v}_i}{dt}$) at frame $t+1$ can be efficiently estimated by subtracting two successive OF vectors:

$$\frac{d\boldsymbol{v}_i}{dt} \simeq \boldsymbol{a}_{\boldsymbol{i}}^{t+1} = \boldsymbol{o}_i^{t+1} - \boldsymbol{o}_i^{t} \tag{5}$$

Estimation of heuristic rule `H2`. According to Eq. 2, the formulation of body contact force involves computing the unit vector $\boldsymbol{n}_{ji}$ for all particles i and j which

[2] According to the physics motion laws.

is not computationally efficient (it is quadratic in the number of particles). It is obvious that body interaction occurs when individual j moves toward individual i and contacts his/her body at time t. This implies that in the case of body contact, the moving direction of j toward i ($\boldsymbol{v}_j$) is similar to the direction of $\boldsymbol{n}_{ji}$ (according to the definition). Furthermore, body contact changes the velocity of individual j, $\boldsymbol{v}_j$, at time t (individual i is considered an obstacle). As a result, $\boldsymbol{n}_{ji}$ can be effectively estimated by acceleration (corresponding to a velocity change) at a very low computational cost. Based on above explanation, we estimate the contact force of particle i caused by its neighboring particles j's as:

$$\boldsymbol{F}_i^{bc} = \frac{\sum_j \boldsymbol{a}_j \cdot \mathbf{g}_i(j)}{\sum_j \mathbf{g}_i(j)} \tag{6}$$

where, $\boldsymbol{a}_j$ is the acceleration vector of particle j (Eq. 5). $\mathbf{g}_i(j)$ is defined by a Gaussian function with bandwidth R as $\mathbf{g}_i(j) = \frac{1}{\pi R^2} \exp\left(\frac{-d_{ij}^2}{R^2}\right)$, where d_{ij} is the Euclidean distance of particles i and j. In practice, Eq. 6 for all particles can be estimated by simply convolving a precomputed 2D Gaussian function over the acceleration map. The magnitude of body contact force, which is the map of `H2`, is referred to as *body compression*.

Estimation of heuristic rule `H3`. To estimate the accumulated aggression force imposed on particle i from its opponent particles, we re-formulate Eq. 3 as:

$$\boldsymbol{F}_i^{agg} = \sum_j \left(\boldsymbol{n}_{ji} \cdot \mathbf{w}_{ij} \cdot \mathbf{f}_{\boldsymbol{o}_i}^{\alpha}(j)\right) \tag{7}$$

where, using OF, the aggression factor $\mathbf{w}_{ij}$ is defined as:

$$\mathbf{w}_{ij} = \frac{1}{2} \cdot \left(1 - \frac{\boldsymbol{o_i} \cdot \boldsymbol{o_j}}{\|\boldsymbol{o_i}\| \cdot \|\boldsymbol{o_j}\|}\right) = \frac{1}{2} \cdot (1 - \cos\phi_{ij}) \tag{8}$$

such that ϕ_{ij} is the angle between the optical flows $\boldsymbol{o}_j$ and $\boldsymbol{o}_i$ of the j^{th} and i^{th} particles, respectively.

Computing the aggression factor $\mathbf{w}_{ij}$ for particle i requires to calculate the cosine between $\boldsymbol{o}_j$ and $\boldsymbol{o}_i$ for each i and j which is quadratic in the number of particles. To reduce the computations, therefore, we propose two approximations of $\mathbf{w}_{ij}$ over Q quantized bins of OF orientations, θ^q, $q = 1, ..., Q$, instead of directly computing them exhaustively. θ_i^q indicates the bin to which the orientation of OF vector $\boldsymbol{o}_i$ (with respect to a fixed reference axis) belongs. As the first approximation, we set $\mathbf{w}_{ij} = 1$ when $\theta_i^q = -\theta_j^q$ and zero otherwise, denoted by $\tilde{\mathbf{w}}_{i,j}^{[1]}$. This implies that the aggressive factor of the particle i depends on its neighboring particles approaching particle i from *exactly* opposite quantized direction. As second approximation, $\mathbf{w}_{ij} = 1$ when the orientations of $\boldsymbol{o}_i$ and $\boldsymbol{o}_j$ do not fall in a same quantized bin, $\theta_i^q \neq \theta_j^q$, and zero otherwise, $\tilde{\mathbf{w}}_{i,j}^{[2]}$. This approximation, on the other hand, states that any particle approaching particle

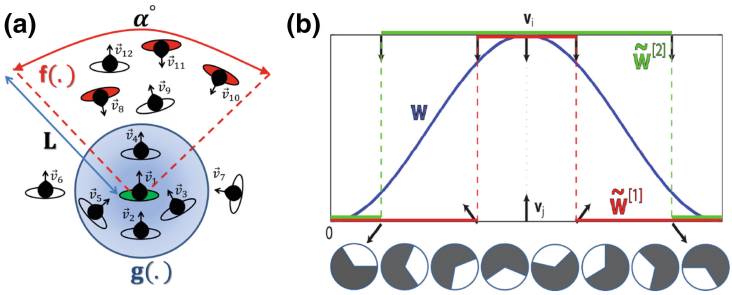

Fig. 2. (a) The windowing function $\mathbf{g}(\cdot)$ returns non zero value for particles inside the circle and zero for the rest. The window function $\mathbf{f}(\cdot)$ simulates the view field of the green particle and returns non-zero for the articles in the view field. The particles marked in red are considered as opponents approaching the green particle. (b) top: two approximations of $\mathbf{w}$ varying the direction of $\boldsymbol{v}_j$ as shown on the bottom x-axis respect to $\boldsymbol{v}_i$ with fixed direction on the top x-axis, Eq. 9. Bottom: the binary filters $\mathbf{f}_q^\alpha$ modeling the particle's view field, $Q = 8$ and $\alpha = 120°$ (Best viewed in color)

i from a different orientation (bin) is considered in the aggression factor. The first and the second approximations, $\tilde{\mathbf{w}}_{i,j}^{[1]}$ and $\tilde{\mathbf{w}}_{i,j}^{[2]}$, are defined as follows:

$$\tilde{\mathbf{w}}_{i,j}^{[1]} = \begin{cases} 1 & \text{if } \theta_i^q = -\theta_j^q \\ 0 & \text{otherwise} \end{cases} \qquad \tilde{\mathbf{w}}_{i,j}^{[2]} = \begin{cases} 1 & \text{if } \theta_i^q \neq \theta_j^q \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

(Figure 2 b-top) illustrates the real values of $\mathbf{w}_{i,j}$ (Eq. 8) and its approximations $\tilde{\mathbf{w}}_{i,j}^{[\cdot]}$ (Eq. 9), where the black arrows indicate the directions of particle i and its neighboring particles j's. It is shown that $\tilde{\mathbf{w}}_{i,j}^{[1]}$ is 1 only for particles approaching i from opposite direction (and zero from other directions), while $\tilde{\mathbf{w}}_{i,j}^{[2]}$ is 1 for particles approaching i from any different direction with respect to i's direction.

According to the heuristic rule `H3`, the windowing function $\mathbf{f}_{\boldsymbol{o}_i}^\alpha(\cdot)$ should reflect what each particle sees (i.e., individual's view field). Therefore, we define it as a naive-shaped that resembles one's field of view, oriented in the direction of the particle's optical flow $\boldsymbol{o}_i$. Here we are making the fair assumption that a pedestrian looks at his/her walking direction, which is especially valid in crowd scenarios. We set the angle of view α to 120° as in human vision. The definition of angle of view α and length of view field L in $\mathbf{f}_{\boldsymbol{o}_i}^\alpha(\cdot)$ is illustrated in Fig. 2(a). In practice, we model particle's field of view on the image plane using a fixed filter bank composed of Q filters (binary masks), $\{\mathbf{f}_q^\alpha\}_{q=1}^Q$, where each filter implies a quantized orientation bin as illustrated in (Fig. 2 b-bottom) for $Q = 8$ and $\alpha = 120°$. Similar to `H2`, computing $\boldsymbol{n}_{ji}$ is also quadratic in the number of particles. According to `H3`, individual i moves towards individual j from the coordinates i to j. This implies that the direction of $\boldsymbol{v}_i$ is in the opposite of $\boldsymbol{n}_{ji}$. Therefore, for the sake of complexity, we approximate $\boldsymbol{n}_{ji} \simeq -\boldsymbol{v}_i$.

Taking into account all the approximations, the aggression force on particle i is estimated as:

$$\boldsymbol{F}_i^{agg} \simeq -\boldsymbol{o}_i \cdot \sum_{q=1}^{Q} \left([\theta_i^q = q] \cdot \left(\tilde{\mathbf{w}}_i^{[\cdot]} \star \mathbf{f}_q^{\alpha} \right)(i) \right) \tag{10}$$

where $[\cdot]$ is a indicator function that returns 1 if $\theta_i^q = q$ and zero otherwise, and $\star$ is the convolution operator which identically performs the summation over neighboring particles j in Eq. 7. $\left(\tilde{\mathbf{w}}_i^{[\cdot]} \star \mathbf{f}_q^{\alpha} \right)(i)$ is the value of the convolution at the coordinates of particle i. According to the Convolution Theorem [28], Eq. 10 can be efficiently computed in the Fourier domain. We called the magnitude of aggression force as *aggressive drive*.

***Visual Information Processing Signature* - VIPS.** Each heuristic captures a different aspect of visual information processed by individual cognition in crowd scenarios. To define a single informative feature, we simply combine together acceleration, body compression and aggression drive in a feature we called Visual Information Processing Signature, in short VIPS.

More specifically, we employ the standard bag-of-words (BOW) paradigm *separately* for each of the three maps (Eqs. 5, 6 and 10). Then, for each video clip we sampled P patches of size $5 \times 5 \times 5$ from locations where the corresponding optical flow is not zero, and we build a visual dictionary of size K using K-means clustering[3]. In the BOW assumption, each video is encoded by a bag; to compute such bags we assign each of the P patches to the closest codebook, and we pool together all the patches to generate an histogram over the K visual words. *The final VIPS is obtained by concatenating the histograms resulting from acceleration, body compression and aggressive drive.* This process is illustrated in the right-most part of Fig. 1.

To address the specific approximations, we employed for the aggressive drive, $\tilde{\mathbf{w}}_{i,j}^{[1]}$ or $\tilde{\mathbf{w}}_{i,j}^{[2]}$ (see Eq. 9), in the experiments we will refer to our descriptor as VIPS$^{[1]}$ and VIPS$^{[2]}$, respectively. Finally, to further validate the aggressive drive, we also considered a third baseline version of $\tilde{\mathbf{w}}$ in which we did not remove any orientation (Eq. 9), and we simply filtered the quantized OF with the wedge filters of (Fig. 2 b-bottom). We will refer to this baseline as VIPS$^{[*]}$.

5 Experiments

We evaluate our approach on three standard benchmarks namely Violence in Crowds (VIC) [18], Violence in Movies (VIM) [20] and BEHAVE [29] datasets. In particular, VIC is the only available dataset specifically assembled for classifying acts of violence in crowd scenes, while VIM allows us to evaluate the robustness of our approach in person-on-person violent scenes. We also select

[3] To employ K-means, we rasterize each patch in a vector of length 125 along with the Euclidean distance. we empirically set $K = 500$ selected from a range of $[100, 200, ..., 2000]$.

Fig. 3. First three columns are frame samples taken from Violence in Crowds (VIC), Violence in Movies (VIM), BEHAVE and Violence-Cross (VC) datasets, respectively. Reader is encouraged to review the text for details.

BEHAVE dataset, which constitutes several complex group activities (e.g., walking together, splitting, escaping, and fighting). Besides, we realized that the most similar behavior to our first approximation ($VIPS^{[1]}$) is "crowd crossing" in which people cross a road in opposite directions. Therefore, to show the robustness of the proposed method to distinguish violent from crossing behaviors in normal situations, we create a new dataset called Violence-Cross (VC) whose videos gathered from VIC dataset and CUHK dataset [30]. It includes 300 videos, equally divided into three classes (100 videos for each class). *Class 1* consists of videos of violent behaviors, *Class 2* contains videos of people walking in opposite directions (cross walk), and *Class 3* contains videos showing actions different than violent and crowd crossing behaviors (e.g., marathon, crowd walking in a same direction). The last column of Fig. 3 shows some sample frames of this new dataset.

Effect of varying filter size. We examined the performance of body compression force (F^{bc}) and aggression force (F^{agg}) with respect to different length of the view field L and filter size (Gaussian bandwidth) R, respectively, on VIC dataset[4]. We set the number of random patches to 1000 and varied R and L as $\beta * max(h, w)$ pixels (for both R and L), where $\beta \in \{0.025, 0.05, 0.075, 0.1, 0.15\}$ and $h \times w$ is the dimension of video frame (320×240 in this case). Figure 4(a) shows that a larger length of field of view results in better performance for aggression forces (F^{agg}). However, we observed that increasing size of the Gaussian filter leads to decreasing performance of the body compression (F^{bc}). This is indeed consistent with our definition of body contact force, where only particles (individuals) that are really close may impose body compression forces.

Effect of number of random sampled patches. We evaluated the performance of VIPS varying P, the number of random patches extracted from each video or clip. We empirically set β for R and L to 0.025 and 0.1, respectively (i.e., $R = 8$ and $L = 32$ pixels). We varied $P \in \{50, 100, 200, 400, 800, 1000\}$. Figure 4(b) summarizes the results. As expected, the accuracy on VIC and VIM are improved by increasing the number of sampled patches P. Interestingly, $VIPS^{[1]}$ outperformed $VIPS^{[2]}$ and $VIPS^{[*]}$ for all the P values on all datasets.

[4] Without concatenating them to form the final VIP signature.

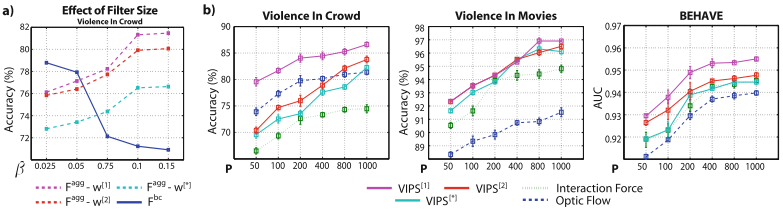

Fig. 4. Evaluating (a) the effect of filter size on aggression and body compression force values, and (b) the effect of varying number of random sampled patches.

This supports our choice of considering individuals approaching from opposite direction as opponents. Finally, the results show the superiority of VIPS compared to optical flow and interaction force (SFM) [3] methods with respect to different number of sampled patches.

Comparison with the state of the art. We compared our approach with the Interaction Force (SFM) [3], Acceleration Measure Vector (AMV) [14], optical flow [27], and ViF [18] as baselines, and some state-of-the-art descriptors used for violent acts from crowd videos including MoSIFT [17,20], and Substantial Derivative (SD) approach [31]. Moreover, in order to demonstrate the effectivness of the proposed method, we compared with ConNet. Although there is no existing pre-trained ConvNet network exist for violence detection, mainly due to scarcity of training example, we evaluate our method with pre-trained model on WWW-crowd dataset [32], which is the most relevant pre-trained model for crowd behavior analysis. We first construct the feature vector by getting the average deep features vector of 10 jittered samples of the original image. Then, we L_2 normalized the feature vectors, and evaluate its performance on VIC, VIM, and BEHAVE datasets. We performed violence classification at video level for VIC, VIM, and VC datasets. For the first two datasets, we followed the standard training-testing splits that come with each dataset, whilst for the VC we equally divide each class into a test set of 150 videos (50 video sequences for each class) and the rest for testing. Then, we compute VIPS for each video and a Support Vector Machine (SVM) with Histogram Intersection Kernel [17] is adopted for video classification. However, for the BEHAVE dataset, the associated task is temporal detection by assigning either normal or abnormal (violent) label to each frame of a video. For this purpose, we computed VIPS at frame level. Since abnormal data is not available in the training time, following the standard procedure of [3], we employed Latent Dirichlet Allocation (LDA) [33] to generatively model normal crowd behaviors. In order to compensate the effect of random sampling, we repeated each experiment 10 times, reporting mean performance. It is also worth mentioning that, for all the experiments, we employed four quantized orientations to compute the aggression force, i.e., $Q = 4$ in Eq. 10. We also tried larger values of Q, but results did not improve. We set filter sizes to $L = 32$ and $R = 8$ and select $P = 1000$ with size of $5 \times 5 \times 5$. Table 1 reports the comparison with the state-of-the-art methods as well as the performance

of each element of VIPS descriptor on VIC, VIM, and BEHAVE datasets. As immediately visible in dense (VIC) and moderate crowd scenes (BEHAVE), the first approximation of aggression force ($F^{agg} - W^{[1]}(H_3^{[1]})$) shows a better performance as compared to acceleration (H_1), Body Compression (H_2), and the second approximation of aggression force ($F^{agg} - W^{[2]}(H_3^{[2]})$). In addition, we observe that for person-to-person violent situations, H_2 and $H_3^{[1]}$ show very similar performance, and their combination with acceleration (VIPS$^{[1]}$) improved the overall performance of the classifier for all scenarios including moderate crowd scene (BEHAVE dataset). However, we can see that, as compared to the Energy Potential descriptor [23], VIPS$^{[1]}$ does not achieve significant improvement. We believe that this is mainly due to our sampling strategies, and that the results can be improved using trajectory-based method. Nonetheless, we conclude that VIPS$^{[1]}$ has a strong discriminative power on detecting violent behaviors regardless the scene crowdedness (from dense to moderate crowd scenes, as well as person-to-person fight). As an example, MoSIFT [17,20] obtained very promising accuracy (the second best after our approach) on VIM (person-to-person), but poor performance on VIC (which is characterized by a dense crowd). This states that, unlike our approach, MoSIFT is sensitive to the crowd density. Moreover, the SFM and AMV obtained very competitive accuracy on VIM, while their performance on VIC drastically decreased. This supports the discussion in the socio-psychology literature [7,25] reporting that social force models perform poorly in overcrowded situations, since they are not capable of modeling complex behavioral patterns in such scenarios. In addition, one can observe that ConNet based approach obtained significant inferior performance compared with hand-crafted competitors on BEHAVE and VIM, however, it gained comparable performance on VIC. This is also understable since VIC has a closer characteristic to the source database used for training the pre-trained network compared with VIM and BEHAVE dataset. Finally, we evaluate robustness of our descriptors to distinguish between acts of violence from crossing behaviors. In particular, we conducted experiments on each element of VIPS to show their contributions to the final performance. Moreover, we select ViF [14] descriptor which was designed for detecting violent behaviors in crowd and SFM [3], which is considered as one of the most well-known descriptor to detect abnormality in crowds. Figure 5 shows the confusion matrices of two state-of-the-art methods and elements of the proposed method. We observe that ViF shows a good performance on detecting acts of violence compared to the SFM, however, its overall accuracy is low since it is much confuse to distinguish violent from normal and crossing behaviors. On the other hand, similar to what we observed in the previous experiments, $H_3^{[1]}$ plays an important role in distinguishing violent behaviors, which results in significantly high performance on $VIPS^{[1]}$, able to well discriminate among the three classes.

Runtime performance. The final experiment evaluated the complexity (runtime) of computing the proposed video signature comparing to the real time violent-flows descriptor [18]. The time for BOW encoding is not considered in

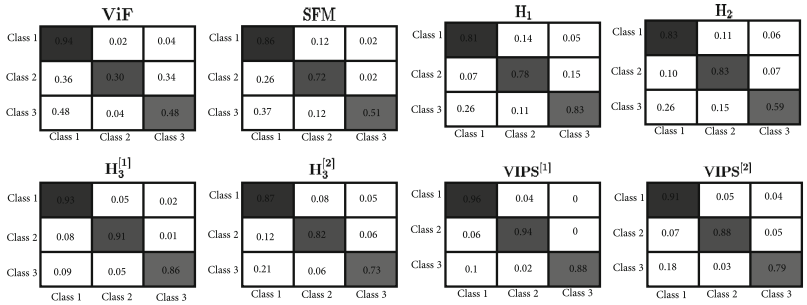

Fig. 5. Average accuracy on Violent-Cross dataset. Class1, Class2, and Class3 are referred to as violent, cross walk, and normal behaviors, respectively. ViF [18] with 57 % overall accuracy; SFM [3] with 69 % overall accuracy, Acceleration (H_1) with 74 % overall accuracy, Body Compression (H_2) with 75 % accuracy; (bottom, first): Aggression force ($H_3^{[1]}$) with 89 % overall accuracy, Aggression force ($H_3^{[2]}$) with 80 % overall accuracy, $VIPS^{[1]}$ with 92 % overall accuracy and $VIPS^{[2]}$ with 86 % overall accuracy.

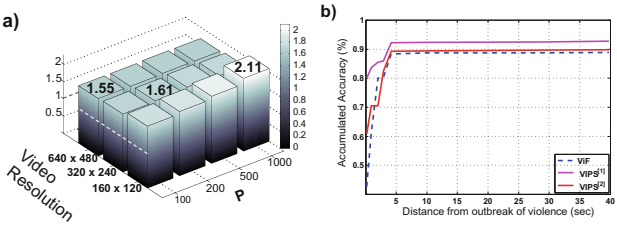

Fig. 6. Runtime performance. (a) Evaluating the running time of VIPS with respect to different sampled patches and video resolutions. (b) Accumulated accuracy on 21 videos as a function of distance from violence outbreak.

this experiment (the real time efficiency of BOW encoding is shown in [34]). First, we measured the relative computational time of our method with respect to violent-flows [18]. Figure 6(a) shows the ratio between time to process a clip for VIPS and violent-flows as a function of number of sampled patches and video resolution. For both methods we employed the same implementation of the optical flow in [27]. The results show that our method is roughly 1.5 to 2 times slower compared to [18]. During the experiments, we observed that the dominant computational cost of our method belongs to the optical flow computation, in particular for medium-to-high resolutions, whereas the convolutions (in the frequency domain) add negligible computational burden. Second, we evaluated the accuracy and detection time of both methods. For this purpose, following [18], we selected 21 videos from the VIC dataset that start with a non-violent behavior and then turn to violent situations mid-way through the video. The goal is to detect the violence as close to its annotated violence start point (outbreak). Figure 6(b) summarizes the results, where our approach (VIPS[1] and

Table 1. Average accuracy over 10 times of repeated trials for the VIC, VIM and BEHAVE datasets.

	Violence in crowds	Violence in movies	BEHAVE
Optical flow [27]	78.48 %	91.31 %	93.48 %
SFM [3]	74.50 %	95.51 %	94.23 %
AMV [14]	74.18 %	95.02 %	86.72 %
SD [31]	85.43 %	96.89 %	94.8 %
ConvNet [32]	83.48 %	89.52 %	79.12 %
MoSIFT [17,20]	83.42 %	89.50 %	-
ViF [18]	81.30 %	-	-
Energy potential [23]	-	-	94.50 %
$Acceleration(H_1)$	79.14 %	93.40 %	90.23 %
$F^{bc}(H_2)$	78.83%	94.10%	92.07%
$F^{agg} - W^{[1]}(H_3^{[1]})$	81.87%	95.12%	91.15%
$F^{agg} - W^{[2]}(H_3^{[2]})$	78.45 %	94.23 %	89.87 %
VIPS$^{[1]}$	**86.61 %**	**96.91 %**	**95.73 %**
VIPS$^{[2]}$	83.77%	96.51%	94.3%
VIPS$^{[*]}$	82.26%	96.11%	94.26%

VIPS$^{[2]}$) obtained higher accumulated accuracy for all the expected detection delays. This test, overall, shows that our approach outperforms ViF with slightly higher computational cost. The curve of ViF is fixed after five seconds meaning that its accuracy is not improved anymore.

6 Conclusions

This paper introduced a novel framework to identify violent behaviors in crowd scenes. In particular, we have proposed three behavioral heuristic rules to model a wide range of complex actions underlying crowd scenarios. We explained how to formulate the behavioral heuristics in computational terms and how to estimate them with very low complexity from video sequences. Experimental results illustrated that the proposed approach is not only computationally efficient, but also it is highly robust to various situations in terms of crowd density and different crowd behaviors, such as crossing and fighting, various imaging conditions, occlusions, and camera motions to name a few. Moreover, we observed that the proposed aggressive drive force has a considerable ability to localize regions of conflict at the pixel level, as compared to other descriptors such as optical flow and SFM. However, due to lack of annotated data, we were not able to fully present this type of evaluation. A potential weakness of this work is using fixed-size filter regardless of the scene properties and imaging conditions, which may have a negative impact on the performance. Both the latter aspects require further investigations and will be subject of future work.

Acknowledgement. We thank Daniele Meneghelli (Trinity Analysis and Investigations, Dublin), for the useful discussion on the behavior of violent crowds that partly inspired our definition of the aggression force.

References

1. Hoffman, K.: Criminological theories: Introduction, evaluation and application. Teach. Sociol. **28**(4), 403 (2000)
2. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
3. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009. IEEE, pp. 935–942 (2009)
4. Zeng, W., Nakamura, H., Chen, P.: A modified social force model for pedestrian behavior simulation at signalized crosswalks. Procedia Soc. Behav. Sci. **138**, 521–530 (2014)
5. Parisi, D.R., Gilman, M., Moldovan, H.: A modification of the social force model can reproduce experimental data of pedestrian flows in normal conditions. Phys. A **388**(17), 3600–3608 (2009)
6. Zanlungo, F., Ikeda, T., Kanda, T.: Social force model with explicit collision prediction. EPL (Europhys. Lett.) **93**(6), 68005 (2011)
7. Moussaïd, M., Helbing, D., Theraulaz, G.: How simple rules determine pedestrian behavior and crowd disasters. Proc. Natl. Acad. Sci. **108**(17), 6884–6888 (2011)
8. Moussaïd, M., Nelson, J.D.: Simple heuristics and the modelling of crowd behaviours. In: Weidmann, U., Kirsch, U., Schreckenberg, M. (eds.) Pedestrian and Evacuation Dynamics 2012, pp. 75–90. Springer, Heidelberg (2014)
9. Moussaid, M., Guillot, E.G., Moreau, M., Fehrenbach, J., Chabiron, O., Lemercier, S., Pettré, J., Appert-Rolland, C., Degond, P., Theraulaz, G.: Traffic instabilities in self-organized pedestrian crowds. PLoS Comput. Biol. **8**(3), e1002442 (2012)
10. Martignon, L., Hoffrage, U.: Fast, frugal, and fit: Simple heuristics for paired comparison. Theor. Decis. **52**(1), 29–71 (2002)
11. Hutchinson, J.M., Gigerenzer, G.: Simple heuristics and rules of thumb: where psychologists and behavioural biologists might meet. Behav. Process. **69**(2), 97–124 (2005)
12. Hertwig, R., Todd, P.M.: More is not always better: the benefits of cognitive limits. In: Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making, pp. 213–231 (2003)
13. Gigerenzer, G., Goldstein, D.G.: Reasoning the fast and frugal way: models of bounded rationality. Psychol. Rev. **103**(4), 650 (1996)
14. Datta, A., Shah, M., da Vitoria Lobo, N.: Person-on-person violence detection in video data. In: ICPR vol. 1, pp. 433–438 (2002)
15. de Souza, F.D.M., Chvez, G.C., do Valle Jr., E.A., de Albuquerque Arajo, A.: Violence detection in video using spatio-temporal features. In: SIBGRAPI 2010, pp. 224–230 (2010)
16. Déniz, O., Serrano, I., Bueno, G., Kim, T.: Fast violence detection in video. In: Proceedings of the 9th International Conference on Computer Vision Theory and Applications, (VISAPP 2014), vol. 2, Lisbon, Portugal, 5–8 January 2014, pp. 478–485, January 2014
17. Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L.: Violent video detection based on MoSIFT feature and sparse coding. In: ICASSP, pp. 3538–3542. IEEE (2014)

18. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1–6. IEEE (2012)
19. Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., Murino, V.: Analyzing tracklets for the detection of abnormal crowd behavior. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 148–155. IEEE (2015)
20. Nievas, E.B., Suarez, O.D., García, G.B., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, Walter (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23678-5_39
21. Solmaz, B., Moore, B.E., Shah, M.: Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 2064–2070 (2012)
22. Mehran, R., Moore, B.E., Shah, M.: A streakline representation of flow in crowded scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6313, pp. 439–452. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15558-1_32
23. Cui, X., Liu, Q., Gao, M., Metaxas, D.N.: Abnormal detection using interaction energy potentials. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3161–3167. IEEE (2011)
24. Raghavendra, R., Bue, A.D., Cristani, M., Murino, V.: Optimizing interaction force for global anomaly detection in crowded scenes. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 136–143. IEEE (2011)
25. Moussad, M., Nelson, J.: Simple heuristics and the modelling of crowd behaviours. In: Weidmann, U., Kirsch, U., Schreckenberg, M. (eds.) Pedestrian and Evacuation Dynamics 2012, pp. 75–90. Springer, Heidelberg (2014)
26. Scheflen, A.E., Ashcraft, N.: Human territories: how we behave in space-time (1976)
27. Liu, C., Freeman, W.T.: A high-quality video denoising algorithm based on reliable motion estimation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6313, pp. 706–719. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15558-1_51
28. Papoulis, A.: The Fourier Integral and its Applications. McGraw-Hill, New York (1962)
29. Blunsden, S., Fisher, R.: The behave video dataset: ground truthed video for multi-person behavior classification
30. Shao, J., Loy, C.C., Wang, X.: Scene-independent group profiling in crowd. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2227–2234. IEEE (2014)
31. Mohammadi, S., Kiani, H., Perina, A., Murino, V.: Violence detection in crowded scenes using substantial derivative. In: 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2015)
32. Shao, J., Kang, K., Loy, C.C., Wang, X.: Deeply learned attributes for crowded scene understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4657–4666. IEEE (2015)
33. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
34. Uijlings, J., Duta, I., Sangineto, E., Sebe, N.: Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. Int. J. Multimedia Inf. Retrieval **1**, 1–12 (2015)