CROWD MOTION MONITORING USING TRACKLET-BASED COMMOTION MEASURE

*Hossein Mousavi** *Moin Nabi** *Hamed Kiani* *Alessandro Perina* *Vittorio Murino*

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

ABSTRACT

Abnormal detection in crowd is a challenging vision task due to the scarcity of real-world training examples and the lack of a clear definition of abnormality. To tackle these challenges, we propose a novel measure to capture the commotion of a crowd motion for the task of abnormality detection in crowd. The unsupervised nature of the proposed measure allows to detect abnormality adaptively (i.e. context dependent) with no training cost. The extensive experiments on three different levels (e.g. pixel, frame and video) show the superiority of the proposed approach compared to the state of the arts.

***Index Terms*—** Video analysis, abnormal detection, motion commotion, tracklets

1. INTRODUCTION

Abnormal behavior detection in highly-crowded environments plays an important role in public surveillance systems, as a result of worldwide urbanization and population growth. Abnormality detection in crowd is challenging put down to the fact that the movements of individuals are usually random and unpredictable, and occlusions caused by overcrowding make the task even more difficult.

Abnormal events often defined as irregular events deviated from normal ones and vice-versa. The intrinsic ambiguity in this chicken-and-egg definition leads to convert the abnormality detection to an ill-posed problem. For example, slowly walking in a subway station is a normal behavior, but it appears as totally abnormal in the rush hours at the same place due to creating stationary obstacles. This observation demands an *in-the-box* viewpoint about the abnormality in which introducing a *context-dependent* irregularity measure seems crucial. For this purpose, the abnormal behaviors in crowded scenes usually appear as crowd *commotion*, so that anomaly detection is in general a problem of detection of crowd commotion [1, 2]. This school of thought investigated a wide range of unsupervised criteria for this purpose and introduced different commotion measures to the literature. It has also been shown that the measure-based (*unsupervised*) methods may outperform supervised methods, because of the subjective nature of annotations as well as small size of training data [2, 3, 4, 5].

Literature review: Mehran et al. [1] introduce a method to detect abnormal behaviors in crowd scenes using a social force model. Similarly in terms of capturing commotion, in [2] a energy-based model approach has been presented for abnormal detection in crowd environments. In [6] force field model has been employed in a hierarchical clustering framework to cluster optical flows and detect abnormality. Lu et al. [7] introduced a correlation-based measure across spatio-temporal video segments extracted by clustering. Authors in [8] selected a region of interest utilizing the motion heat map and measured the abnormality as the entropy of the frames.

Overview: In this paper, we introduce an unsupervised context-dependent statistical commotion measure and an efficient way to compute it, to detect and localize abnormal behaviors in crowded scenes. For this purpose, the scene of interest is modeled as moving particles turned out from a Tracklet algorithm, which can be viewed as motion field distributed densely over foreground. The particles are grouped into a set of *motion patterns* (prototypes) according to their motion magnitude and orientation, and a *tracklet binary code* is established to figure out how the particles are distributed over the prototypes. Here, a novel statistical *commotion measure* is computed from the binary code for each video clip to characterize the commotion degree of crowd motion.

Contributions: Most of the related works only pay attention to either short-period motion information (e.g optical flow) or long-term observation (e.g. trajectory), we instead employed tracklet as an intermediate-level motion representation. The closest recent work to us is HOT [9, 10]. These two works, however, are different not only because we are unsupervised but also introducing the efficient tracklet assignment method employing binary representation of the motion along with a hash function. We specifically shorten the major contributions of this work as following. **First**, we propose Motion Pattern to represent the statistics of a tracklet at each frame in terms of magnitude and orientation. **Second**,

* Authors contributed equally.

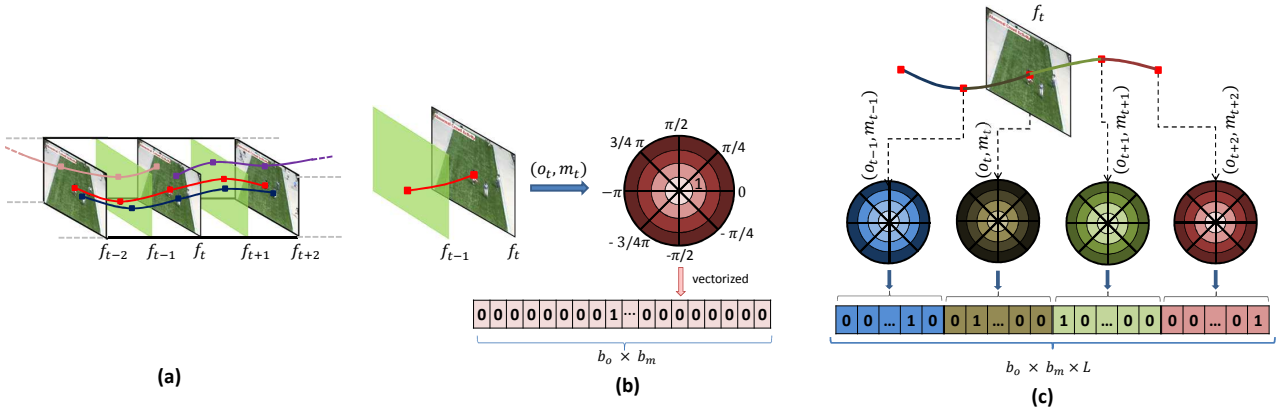

Fig. 1. (a) Four tracklets extracted from corresponding salient points tracked over five frames. (b) A polar histogram of magnitude and orientation of a salient point at the t-th frame of a tracklet (motion pattern). (c) Tracklet binary code is constructed by concatenating a set of motion patterns corresponding salient point over $L+1$ frames.

we propose Tracklet Binary Code representation to model the movement of a salient point over its corresponding tracklet in both spatial and temporal spaces. **Third**, we introduce a new unsupervised measure to evaluate the commotion of a crowd scene in pixel, frame and video levels.

2. THE PROPOSED FRAMEWORK

In this section, we explain the proposed pipeline to measure abnormality of a given video $\mathbf{v} = \{f_t\}_{t=1}^{T}$ with T frames.

Tracklet Extraction: The first step involves extracting all tracklets of length $L+1$ in video $\mathbf{v}$. Towards this purpose, SIFT algorithm is first applied to detect salient points at each frame f_t [11]. Then a tracking technique (we employed the KLT algorithm [12]) is used to track each salient point over $L+1$ frames. Tracklets whose length is less than $L+1$ are considered as noise and, thus, eliminated. The output is a set of tracklets $\mathcal{T} = \{tr^n\}_{n=1}^{N}$, where N is the number of extracted tracklets and tr^n refers to the n-th tracklet. Fig. 1 (a) illustrates a video example and four tracklets which are computed by tracking a set of corresponding salient points over a sequence of five frames.

Motion Pattern: Each tracklet tr^n is characterized by a set of spatial coordinates of its corresponded salient point tracked over $L+1$ frames $\{(x_l^n, y_l^n)\}_{l=1}^{L+1}$. These spatial coordinates are employed to compute the motion orientation and magnitude of the salient point at l-th frame as:

$$o_l^n = \arctan \frac{(y_l^n - y_{l-1}^n)}{(x_l^n - x_{l-1}^n)} \quad (1)$$

$$m_l^n = \sqrt{(x_l^n - x_{l-1}^n)^2 + (y_l^n - y_{l-1}^n)^2} \quad (2)$$

where $2 \leq l \leq L+1$. This step computes a temporary ordered set of L orientations and magnitudes of the salient point corresponded to n-th tracklet $\{(o_l^n, m_l^n)\}_{l=1}^{L}$ (we reset $l=1$ for simplicity).

The motion orientations and magnitudes $\{(o_l^n, m_l^n)\}_{l=1}^{L}$ are used to form a histogram representation of the n-th tracklet. First, a polar histogram h_l^n is computed using the orientation and magnitude of the n-th tracklet at frame l, (o_l^n, m_l^n). This can be easily done by a simple hashing function in $\mathcal{O}(1)$ whose input is (o_l^n, m_l^n) and returns a binary polar histogram with only one "1" value at sector (o_l^n, m_l^n) and zeros for the rest. The polar histogram then is vectorized to a vector of length $b_o \times b_m$, where b_o and b_m are respectively the number of quantized bins for magnitude and orientation. This is illustrated in Fig. 1 (b). The color spectrum of each sector indicates the quantized bin of magnitude. Each arc represents the quantized bin of orientation. We called each vectorized h_l^n a motion pattern.

Tracklet Binray Code: Given a set of orientations and magnitudes, $\{(o_l^n, m_l^n)\}_{l=1}^{L}$, we can correspondingly compute L motion patterns $\{h_l^n\}_{l=1}^{L}$ for the n-th tracklet. Finally, all the (vectorized) motion patterns $\{h_l^n\}_{l=1}^{L}$ are concatenated to compute a tracklet histogram $H^n = [h_1^n, ..., h_L^n]^\top$ of length $b_o \times b_m \times L$ ($\top$ is transpose operator). H is referred to as tracklet binary code, Fig. 1 (c).

Commotion Measuring: To compute commotion measure, each frame f_t is divided into a set of non-overlapped patches $\{p_i^t\}$, where i indexes the i-th patch in the t-th frame. For each patch p_i^t, a subset of tracklet binary codes is selected from $\{H^n\}_{n=1}^{N}$ whose corresponding tracklets spatially pass from patch p_i^t, and p_i^t is temporally located at the middle of the selected tracklets (i.e. if the length of a tracklet is $L+1$, tracklets which start/end $L/2$ frames before/after frame t passing from patch p_i^t are selected). Suppose that N_p tracklet motion codes are selected for patch p_i^t denoted by $\{H^{n_p}\}_{n_p=1}^{N_p}$. Then, we statistically compute the aggregated

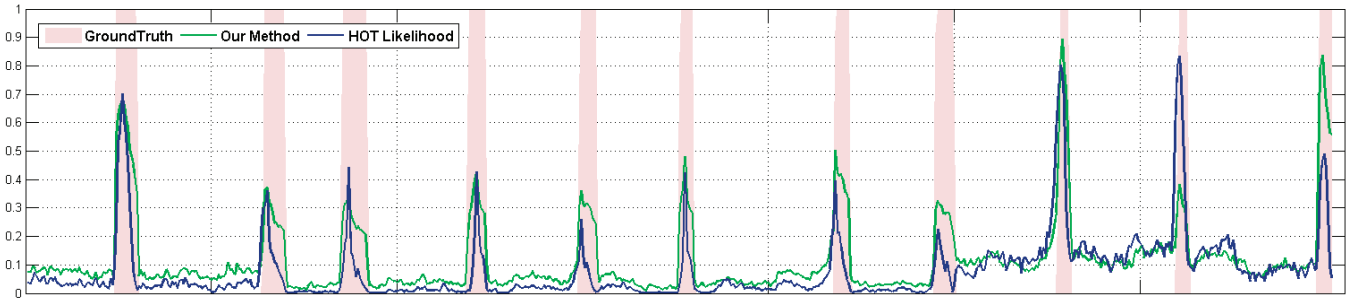

Fig. 2. Results on UMN dataset. The blue and green signals respectively show the commotion measure computed by our approach and LDA+HOT over frames of 11 video sequences. The pink columns indicate the abnormal frames of each sequence. Each sequence starts with normal frames and ends with abnormal frames.

tracklet binary code for patch p_i^t as $\mathcal{H}_i^t = \sum_{n_p=1}^{N_p} H^{n_p}$. The aggregated histogram $\mathcal{H}_i^t$, which contains the distribution of motion patterns of p_i^t, is used to compute the commotion assigned to patch p_i^t as:

$$Comm(p_i^t) = \sum_{j=1}^{|\mathcal{H}_i^t|} w(j, j_{max}) \times ||\mathcal{H}_i^t(j) - \mathcal{H}_i^t(j_{max})||_2^2 \quad (3)$$

where $|.|$ returns the length of vector and j_{max} indicates the index of maximum value in $\mathcal{H}_i^t$ (i.e. $\mathcal{H}_i^t(j_{max})$ is the maximum value in $\mathcal{H}_i^t$). $||.||_2$ is the L2-norm. As mentioned earlier, $\mathcal{H}$ captures the motion patterns distribution of tracklets passing from a sampled patch. As a result, the maximum value of $\mathcal{H}$ indicates the dominant motion pattern over the patch of interest. The amount of commotion of a patch, therefore, can be measured by the difference (deviation) between the occurrences of the dominant motion pattern and the other motion patterns wighted by $w(j, j_{max})$.

$w(j, j_{max}) \in [0, ..., 1]$ is a scalar weight which controls the influence of the j-th motion pattern on the commotion measure respect to the dominant (j_{max}-th) motion pattern. The motivation behind using w is to assign higher(lower) weights to pattern which are less(more) similar to the dominant pattern. The weight $w(j, j_{max})$ is defined using a two-variants Gaussian function as:

$$w(j, j_{max}) = \frac{1}{2\pi\sigma_o\sigma_m} e^{-\frac{(\bar{o}_j - \bar{o}_{j_{max}})^2}{2\sigma_o^2} - \frac{(\bar{m}_j - \bar{m}_{j_{max}})^2}{2\sigma_m^2}} \quad (4)$$

where $\bar{o}_j$ is the middle of the orientation bin that the j-th motion pattern belongs to. For example, if the j-th motion pattern falls in $[0 - \pi/4]$, then $\bar{o}_j$ is $\pi/8$. Similarly, $\bar{m}_j$ is the middle of the magnitude bin that the j-th motion pattern falls in (e.g. if the j-th motion pattern falls in $[3 - 6]$, then $\bar{m}_j$ is $9/2$). The definitions can be identically apply for $\bar{o}_{j_{max}}$ and $\bar{m}_{j_{max}}$. The values of σ_o and σ_m are set to $1/b_o$ and $1/b_m$.

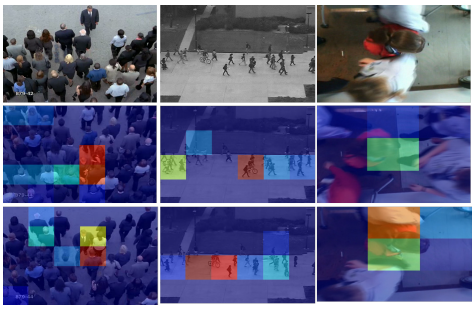

Fig. 3. Qualitative results on sample sequences selected from SocialForce, UCSD and York datasets. The commotion measure of each patch is represented by a heat-map.

3. EXPERIMENTS

In this section, we validate the proposed method in all three settings for abnormality detection including (i) *pixel-level*, (ii) *frame-level* and (iii) *video-level*.

Pixel-level: In this experiment, we evaluated our approach qualitatively on a subset of video sequences selected from standard datasets (UCSD [13], SocialForce[1] and York [14]). For a given video, we first extract a set of tracklets of length $L = 10$. The magnitude and orientation bins are set to 5 and 6 respectively to form the polar histogram (motion pattern). Then each frame is divided to a set of non-overlapped patches in which for each patch the commotion measure is computed using Eq. 3. Qualitative results are shown in Fig. 3 in terms of heat-map respect to the locally computed commotion measure. The selected sequences characterized by different magnitude and orientation, camera view points, different type of moving objects (e.g. human and bike) over scarce-to-dense crowded scenarios. As illustrated in Fig 3 the proposed measure can be effectively exploited for abnormality localization along with a per-defined threshold. Furthermore, the commotion measure can be used as spatial-

temporal interest point for video level abnormality detection (details in video level experiment).

Frame-level: The goal of this experiment is to compute a single commotion measure for each frame of a given video. Toward this purpose, we modified the procedure of computing the commotion measure in three ways: First, each whole frame is considered as a single patch (there is not frame patching). Second a commotion measure is computed for each tracklet passing over the frame of interest. Finally, The measures computed from all the tracklets are summed up as the frame's commotion measure. We evaluated our method on UMN dataset [1] including 11 sequences filmed in 3 different scenes. Fig. 2 shows commotion measure computed for each frame and illustrated as a signal (Green). We compare our method with LDA log-likelihood on HOT, selected as a baseline measure (blue). Since the HOT approach is a supervised technique, unlike the new approach which is unsupervised, these two techniques are not directly comparable. Thus, we divided the dataset into two subsets of video sequences (e.g. A and B). We performed training and testing two times, at each time, a subset (A or B) is selected for training and the other one (B or A) for testing. The LDA log-likelihood for each frame was considered as its commotion measure. Obviously, both approaches perform well and assign lower(higher) commotion measures to normal(abnormal) frames. The difference, however, is that our approach is unsupervised. This is an supreme characteristic for the task of abnormal detection where in most cases there is not a clear definition of abnormality and gathering real-world training videos are intractable. We also obtained the scene-based ROC of our proposed method illustrated in Fig. 4 and the overall Area Under ROC (AUC) in Table 1 comparing with the the leading existing approaches. According to Table 1, our approach achieved the superior detection speed (in terms of frame per second) with very competitive detection performance.

Video-level: In this experiment, we show the effectiveness of the proposed measure when employed as spatio-temporal interest point detector along with one of the best performing descriptors (HOT). In this setting, we adopt the commotion measure to be used in a video-level abnormality detection on Violence-in-Crowd dataset [15]. For this purpose, we first apply a spatio-temporal grid on the video, then for each 3D cell of the grid we compute our proposed measure. For evaluation, we deployed the standard BOW representation pipeline used in most video level works [9, 3]. We enriched the standard setting with a weight vector comes from commotion measure. We simply defined a weight vector with the same length of codebook's size. The weight of each codeword is computed as summation over the commotion measures of the 3D cells belong to its corresponding cluster. Our result outperformed all previous methods including HOT as reported in Table 2.

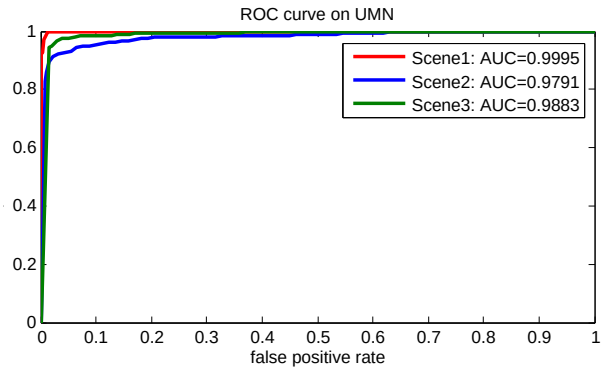

Fig. 4. AUC on 3 scenes of UMN.

Method	AUC	Speed (fps)
Optical Flow [1]	0.84	5
SFM[1]	0.96	3
Chaotic Invariants [16]	0.99	1
Sparse Reconstruction [17]	0.978	<1
Proposed Scheme	0.9889	5

Table 1. Performance of the proposed scheme on UMN.

4. CONCLUSION

The problem of abnormality detection in crowd scenes was addressed in this paper. We proposed a new measure to compute the commotion of a given video and we showed that the new measure can be effectively exploited to detect/localize abnormal events in crowded scenarios. The qualitative and quantitative results on the standard dataset show that our approach outperformed the state of the arts in terms of detection speed and performance. The future direction involves quantitatively evaluate the new approach on more challenging datasets in the pixel level. Moreover, exploring of the proposed approach for the task of action recognition would be a potential direction.

Method	Accuracy
Local Trinary Patterns [18]	71,53%
Histogram of oriented Gradients [19]	57,43%
Histogram of oriented Optic-Flow [20]	58,53%
HNF [19]	56,52%
Violence Flows ViF [15]	81,30 %
Dense Trajectories [21]	78,21 %
HOT [9]	78,30%
Our Method	**81.55**%

Table 2. Classification results on crowd violence dataset, using linear SVM in 5-folds cross validation.

5. REFERENCES

[1] Ramin Mehran, Alexis Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model.," in *CVPR 2009*.

[2] Xinyu Wu Yen-Lun Chen Yongsheng Ou Yangsheng Xu Guogang Xiong, Jun Cheng, "An energy model approach to people counting for abnormal crowd behavior detection," in *Neurocomputing 2012*.

[3] Angela A Sodemann, Matthew P Ross, and Brett J Borghetti, "A review of anomaly detection in automated surveillance," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 1257–1272, 2012.

[4] Tao Xiang and Shaogang Gong, "Video behavior profiling for anomaly detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 5, pp. 893–908, 2008.

[5] M. Nabi, A. Del Bue, and V. Murino, "Temporal poselets for collective activity detection and recognition," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, Dec 2013, pp. 500–507.

[6] P. Huang D. Chen, "Dynamic human crowd modeling and its application to anomalous events detection," in *ICME 2010*.

[7] H. Dongjian L. Yong, "Video-based detection of abnormal behavior in the examination room," in *IFITA 2010*.

[8] C. Djeraba M. Sharif, N. Ihaddadene, "Crowd behaviour monitoring on the escalator exits," in *ICCIT 2008*.

[9] Hossein Mousavi, Ssadegh Mohammadi, Alessandro Perina, Ryad Chellali, and Vittorio Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 148–155.

[10] Hossein Mousavi, Moin Nabi, Hamed Kiani Galogahi, Alessandro Perina, and Vittorio Murino, "Abnormality detection with improved histogram of oriented tracklets," in *International Conference on Image Analysis and Processing (ICIAP)*, 2015.

[11] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] C. Tomasi and T. Kanade, "Detection and tracking of point features," 1991, Intl Journal of Computer Vision.

[13] V. Bhalodia V. Mahadevan W. Li and N. Vasconcelos, "Anomaly detection in crowded scenes," 2010, CVPR.

[14] Andrei Zaharescu and Richard Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Computer Vision–ECCV 2010*, pp. 563–576. Springer, 2010.

[15] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior.," in *CVPR Workshops*. 2012, pp. 1–6, IEEE.

[16] Shandong Wu, Brian E Moore, and Mubarak Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2054–2060.

[17] Yang Cong, Junsong Yuan, and Ji Liu, "Sparse reconstruction cost for abnormal event detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3449–3456.

[18] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 492–497.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, June 2008.

[20] Tian Wang and Hichem Snoussi, "Histograms of optical flow orientation for visual abnormal events detection," in *AVSS*, Washington, DC, USA, 2012, pp. 13–18, IEEE Computer Society.

[21] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, June 2011, pp. 3169–3176.