Unknown Crowd Event Detection from Phase-Based Statistics

Unknown Crowd Event Detection from Phase-Based Statistics

Conference Paper · November 2018

DOI: 10.1109/AVSS.2018.8639174

Alexia Briassouli
Maastricht University

Unknown Crowd Event Detection from Phase-Based Statistics

Alexia Briassouli,
Maastricht University
Bouillonstraat 8-10 6211 LH Maastricht
alexia.briassouli@maastrichtuniversity.nl

Abstract

A new approach for unknown event detection in videos with dense motion, such as crowds or dynamic textures, is developed, without requiring the estimation of optical flow, with no prior knowledge about normal or abnormal events, and with no training. The proposed method directly extracts motion statistics from the phase of the video's Fourier transform and detects changes in them, and in the video, by applying sequential statistical change detection theory. Focus is placed on the motion component, as videos of densely moving entities, such as temporal textures and crowds, often have a very similar appearance, but different dynamic features. Experiments with synthetically generated datasets demonstrate the method's operation under various conditions, while experiments on a recently introduced crowd dataset show that it succeeds at detecting new events in videos of crowds, with no training, and no prior knowledge of the location of new events in space and time.[1]

1. Introduction

Videos with dense motion, such as crowds, traffic, temporal textures, are common in surveillance and monitoring, in security, environmental, commercial and other applications. The detection of unknown events in such videos poses a challenge, due to the lack of knowledge on the nature of the event, the complex nature of the motion in them, occlusions, non-rigidity, and the lack of prior knowledge about the spatiotemporal location of the events.

Many works use motion estimates, tracking or trajectory extraction to detect abnormal events, as motion can provide crucial information about the events taking place, which cannot always be derived from appearance information, especially in crowded scenes with a homogeneous appearance. In [17], crowd motions are modeled as mixtures of dynamic textures for spatiotemporal localization of events, while in [5] tracklets are analyzed to detect abnormal events. Latent topic analysis has also been used for unsupervised event detection in crowds [16], but it requires processing of large datasets, so it has a high computational load, and its outcomes are tailored to the specific dataset being analyzed [14]. Model-based approaches like Linear Dynamic Systems (LDS) [2], [15], or physics-based models [1], [12], cannot sufficiently describe all categories of crowded scene dynamics, which may go beyond each model's restrictions.

Deep learning methods face challenges dealing with the analysis of crowd videos for unusual event detection because abnormal events, by their definition, are not well-defined, can be varied, and are scarce in datasets. This results in a small number of training samples, which tends to lead to overfitting in deep learning approaches. Moreover, as anomalous events are not clearly defined, ground truth annotations in benchmark videos are subjective. A recent deep learning approach for crowd event detection models normal events on benchmark datasets using Generative Adversarial Nets (GANs) and uses them as a basis for the detection of abnormal events, with good results. However, their models are based on optical flow estimation, which can become problematic in very dense scenarios and whose estimation is computationally costly. Slicing CNNs have been proposed in [11] to effectively model crowd motions with a lower computational burden than the very high one of 3D CNNs, and higher accuracy than methods that attempt to represent motion using 2D CNNs on spatiotemporal xt and yt slices of a video in (x, y, t) space. The latter methods contain multiple object motions in each slice, which [11] separate by integrating CNN learned xy features in their framework. This results in improved crowd classification accuracy, but at a higher computational cost, while the issue of abnormal crowd event detection is not addressed. In [9], CNNs pre-trained on image datasets are used to mitigate the heavy computational cost of deep learning, and their temporal variations are fused with optical flow estimates to detect abnormal events in video.

For the analysis of videos with dense motions, such as crowds or dynamic textures, many limitations of existing

[1] Copyright: 978-1-5386-9294-3/18/$31.00 ©2018 IEEE

approaches, including the need to estimate optical flow values, can be overcome by transform domain modeling. In the transform domain each frame is processed globally, so issues related to local sources of noise are overcome, while the computational cost of the methods remains low. Motivated by their use for static textures and the different kinds of spatio-temporal dynamics present in the video data, scale-space transformations have been used [3], [13] to analyze dynamic textures. They are able to capture and separate local and global periodicities in the motions, while wavelets that are more tuned to motion are being developed [3]. Other approaches focus on Fourier-domain processing of the data [4], where phase information of the 2D FT is used for modeling of temporal textures, as it contains most information about the motion in the video sequence. That method differs from ours as it directly uses the phase itself, whereas in this work, we use the FT to extract the characteristic function of stochastic dsiplacements, which provides a complete estimation of motion statistics.

This paper is structured as follows: Sec 2 presents the phase-based approximation of motion statistics, while Sec. 3 details the sequential approach to detecting new events in the video data. The approach is tested experimentally in Sec 4. In Sec. 4.1 there is an in-depth analysis of its properties through testing on synthetically generated datasets, while Sec. 4.2 presents experiments on a newly introduced dataset with crowd motions [8], where its ability to detect new events in crowds is demonstrated.

2. Phase-Based Dense Motion Statistics

In the videos examined in this work, new events are characterized by changes in motion, while the scene appearance often remains the same. For this reason, we introduce a novel approach for the estimation of motion scene statistics, to effectively assess their dynamics and evolution over time. The motion in such videos is complex, and characterized by multiple occlusions, making traditional motion estimation (e.g. by optical flow or other approaches) challenging and potentially inaccurate. For this reason, we make the assumption that multiple small motions can be considered to follow a stochastic distribution which can characterize complex scenes with dense motions more accurately than traditional optical flow or other motion estimates. In the sequel we consider one type of stochastic motion per video frame, however the same framework can be applied locally to small regions of each frame (for example blocks or superpixels).

We consider that there is a displacement $\bar{r} = (r(x,t), r(y,t))$ between frames 1 and t, which results in the Fourier Transform:

$$C(u,v,t) = C(u,v) \cdot e^{-j(ur(x,t)+vr(y,t))}. \quad (1)$$

In our case of (approximately) stochastic motion, we can break down $\bar{r}$ into a deterministic mean $\bar{d}$ (t is omitted for simplicity of notation), and a stochastic zero-mean component $\bar{r}_0$ that follows a distribution $f(\bar{r})$. If all image pixels are displaced by $\bar{r} = \bar{d} + \bar{r}_0$, where $\bar{d} = (d(x), d(y))$ and $\bar{r}_0 = (r_0(x), r_0(y)) \sim f(\bar{r})$, the ratio of the FT of frames 1 and t can be written as:

$$L(u,v,t) = e^{-j(ud(x)+vd(y))} \cdot e^{-j(ur_0(x)+vr_0(y))}, \quad (2)$$

and its 2 D inverse *spatial* FT for random displacement against a zero background is then given by:

$$L(x,y,t) = \delta(x - d(x) - r_0(x), y - d(y) - r_0(y)). \quad (3)$$

In the case of detereministic motion, eq. (3) becomes:

$$L(x,y,t) = \delta(x - d(x), y - d(y)), \quad (4)$$

leading to a single peak around the displacement $\bar{d} = (d(x), d(y))$, which allows its estimation. In the case of stochastic motion, we deduce from eq. (3) that there will be not be one central peak, but a “cloud” of peaks around it, varying with each random motion instantiation.

However, we can derive a comprehensive characterization of the motion statistics from eq. (3). We consider a set of instantiations of the random process represented by $\bar{r}_0$, where, for simplicity, we make the assumption that $\bar{d} = 0$ in the rest of the paper. This is without loss of generality, because a non-zero mean random displacement $\bar{r}$ will simply follow a shifted distribution $f(\bar{d}+\bar{r})$ instead of $f(\bar{r})$. Thus, for zero-mean stochastic displacement, eq. (2) becomes:

$$L(u,v,t) = e^{-j(ur_0(x)+vr_0(y))}. \quad (5)$$

If we consider several instantiations of $\bar{r}_0$, the ensemble average of eq. (5) is given by:

$$E[L(u,v,t)] = E[e^{-jur_0(x)}]E[e^{-jvr_0(y)}], \quad (6)$$

where we have made the simplifying assumption that the motion components in the x and y directions are independent from each other. It is known from probability theory that the characteristic function of a random variable Z that follows the real probability density function (pdf) $f(z)$ is given by its pdf's Fourier transform (FT) (or the complex conjugate of the FT, depending on the definitions used for the characteristic function and the FT):

$$\Phi_Z(v) = \mathcal{F}[f(z)] = \int_{-\infty}^{+\infty} f(z)e^{jvz}dz = E[e^{jvz}], \quad (7)$$

so eq. (6) can be expressed as:

$$E[L(u,v,t)] = \Phi_{r_0(x)}(u)\Phi_{r_0(y)}(v), \quad (8)$$

where $\Phi_{r_0(x)}(u)$, $\Phi_{r_0(y)}(v)$ are the characteristic functions of the random displacements $r_0(x)$ and $r_0(y)$ in the x and y

directions respectively. The characteristic function offers a complete description of a random variable's distribution [7], as all existing higher order moments and the pdf of the random variable can be derived from it. We examine the x direction, and the same procedure can be applied to the y direction. For $v = 0$, eq. (8) becomes:

$$E[L(u, 0, t)] = \Phi_{r_0(x)}(u)\Phi_{r_0(y)}(0) = \Phi_{r_0(x)}(u), \quad (9)$$

since $\Phi_{r_0(y)}(0) = E[e^{j0}] = 1$. Then:

$$f_{r_0}(x) = \mathcal{F}^{-1}[\Phi_{r_0(x)}(u)],\ f_{r_0}(y) = \mathcal{F}^{-1}[\Phi_{r_0(y)}(v)]. \quad (10)$$

i.e. the random displacement's pdf is estimated from the inverse FT of its characteristic function. Eq. (10) can be used to extract the pdf's of the random displacements in the x and y directions using only the FT's of frames 1 and t that were used to estimate $L(u, v, t)$.

In theory, the ensemble average of $L(u, v, t)$ can be estimated from several instantiations of the random process under examination (in our case the random displacements $\bar{r}_0$). In practice this is not possible since we only have one instantiation of the random process of interest, namely one video sequence. To overcome this issue, we make the assumption that the random displacements' process is weakly ergodic, so its statistical properties can be approximated by arithmetic means rather than ensemble averages. Then, in order to extract statistical properties of the random motions, we consider that instantiations of the same $\bar{r}_0$ take place in w_0 frames before and after frame t, giving the arithmetic mean approximation of the ensemble average:

$$E[L(u, v, t)] = \frac{1}{w_0} \sum_{k=t-w_0/2+1}^{t+w_0/2} L(u, v, k). \quad (11)$$

As the experimental results show, this approximation of the ensemble average works well in practice, giving reliable change detection and recognition results. The value of w_0 is set equal to 20, which is long enough for numerical accuracy, but short enough to contain stable motion characteristics (i.e. it is not likely that there will be a significant change in motion during those frames).

3. Sequential Cumulative Sum for Change Detection

When a new event occurs in videos of crowds or traffic, it often affects the motion in the scene more significantly that its appearance. This occurs, for example, in videos of crowds of people who are walking and suddenly run and/or disperse, or in traffic videos where the traffic may change from light to heavy. The pdfs of the displacements estimated as in Sec. 2 can then be used in a statistical sequential change detection framework in order to detect changes in the scene motion. Sequential change detection methods are used to detect changes in a dataset's statistical distribution (in our case the pdf of the displacements) using only currently available data, and can therefore be implemented for real time solutions. In this work we focus on the Cumulative Sum (CUSUM) algorithm, as it is designed to sequentially detect changes between distributions.

If the data under examination at time t is $X_t = [x_1, x_2, ..., x_t]$, its distribution before an unknown change point k^* is $f_0(x)$, and after this change point it is $f_1(x)$, the log-likelihood ratio at each time instant k is given by:

$$s_k = \ln\left(\frac{f_1(X_k)}{f_0(X_k)}\right) \quad (12)$$

and the CUSUM test statistic is given by [6]:

$$T_k = (T_{k-1} + s_k)^+, \quad (13)$$

where $(\cdot)^+ = max(0, \cdot)$. A change is detected when the test statistic T_k surpasses a threshold η, which is chosen using training data to simultaneously lead to the smallest number of false alarms and the quickest detection of changes.

4. Experiments

We have carried out experiments on synthetic "dynamic texture" videos, whose texture has a homogeneous appearance and stochastic motion that changes over time, to simulate real-world crowd/traffic data and dynamic textures, as well as with real-world crowd datasets. Our experiments demonstrate that the proposed method detects changes quickly, with no prior knowledge on the type of video being analyzed, and no training, as well as no estimation of optical flow.

4.1. Synthetic Dynamic Textures

In order to demonstrate the effectiveness of the proposed approach, we first create synthetic videos of dynamic textures, that simulate new events in real-world videos. The synthetic videos feature a homogeneous textured appearance undergoing stochastic motion that changes with time, either over the entire frame, or over one part of the frame with the same appearance. The homogeneity of the moving texture's appearance shows that the results cannot be attributed to its changing appearance, but only to the motion distribution estimated by our approach. Fig. 1 shows two frames of the textured video, comprising of blood cells, and the absolute difference between successive frames, to demonstrate the warping of all frame pixels by the stochastically generated two-dimensional displacement. The difference values are very low, so they are magnified by 255 in the figure, showing that all frame pixels are warped by different values of the normal distribution.

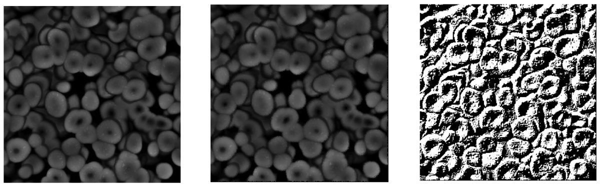

Figure 1. Synthetic stochastically moving texture video: frame 1, 10, difference of frames 10 and 1 magnified by 255 for visibility.

In the first set of experiments, the video frames undergo stochastic motion by warping all frame pixels by x and y displacements following a normal distribution with $\mu = 1$ and $\sigma = 0.5$. The initial motion distribution is approximated through Eq. (11), where the averaging of the Fourier transform ratio takes place over the first w_0 synthetic video frames. These experiments examine the effect of varying values of w_0 on the accuracy of the change detection, by comparing the resulting CUSUM curves. Fig. 2 shows that the change at $N/2$ is clearly detected for all values of w_0 examined, but $w_0 = N/20$ was chosen as achieving a more clearly defined "elbow" in the CUSUM curve, making change detection more robust.

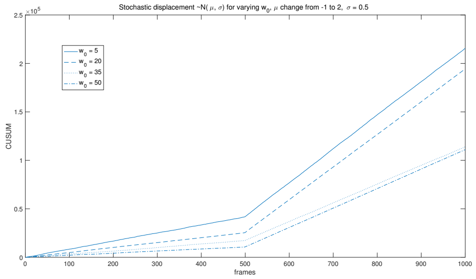

Figure 2. CUSUM values for varying w_0

In the next set of experiments, we keep $w_0 = 20$, set $\mu = 1$, $\sigma = 0.3$ for the initial distribution, and at frame $N/2 = 500$ we change the motion distribution's variance. We examine the CUSUM curves for $\sigma = 0.1 : 2$ to see how it behaves for a number of stochastically displaced textures. The results in Fig. 3 show that, as before, the change is clearly detected when it takes place.

We also examine the effect of varying stochastic displacement by fixing $w_0 = 20$, $\sigma = 0.3$ and varying the mean $\mu = -4 : 4$. In Fig. 4 we see that the CUSUM curve still identifies the change-point reliably, despite the relatively high variance of the displacement values, in relation to their mean values.

4.2. Motion-Emotion Dataset

The Motion-Emotion Dataset (MED) [8] has been recently released for the detection of crowd events which are related to emotions, with the categories: Panic, Fight, Congestion, Obstacle, Neutral. It is suited for testing our method's change detection, as ground truth is provided for the different emotion-relevant crowd events. The crowds in it are of moderate density, but feature frequent occlusions, which would create errors in optical-flow estimates. It should be noted that the categories Neutral and Nothing are essentially the same event, since the motion of the people in the scene remains the same in those cases.

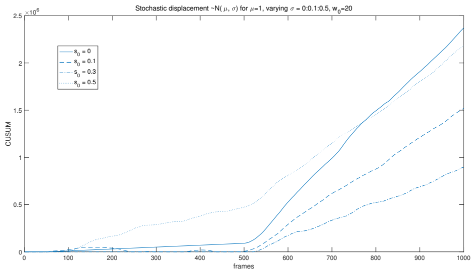

Figure 3. CUSUM values for varying σ

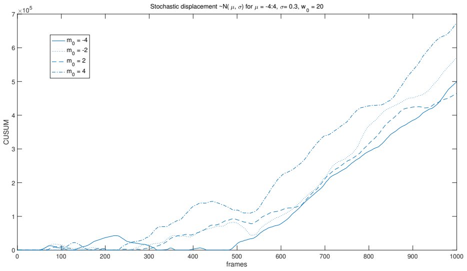

Figure 4. CUSUM values for varying μ

We carried out experiments using the ground truth provided for this dataset, however, upon its careful examination we observed different start/end times of events. This in part due to the transitory nature of the event occurrences, and in part to the subjective nature of such annotations. In our experimental results, the changes detected closely correspond to changes in the crowd density itself, but are also close to the provided ground truth. Table 1 presents the experimental results, where it can be seen that the changes are detected quite accurately, despite the lack of training. The average error, normalized per the number of frames in each video, was calculated to be equal to 0.118, however this is taking into account false alarms and deviations from the ground truth that correspond to actual changes in motion. Most importantly, our comparisons are only with the provided ground truth and not with the methods described in [8], [10], as the latter are detecting changes based on machine learning based classification results, instead of direct change detection. In Fig. 5 we depict frames from video 7, where the crowd avoids an obstacle and bursts into panic. When avoiding the obstacle, the crowd motion flow changes, while the frames where panic takes place are very shaky and the individuals are running in all directions. As a result, the CUSUM, depicted in Fig. 6 displays a sharp change near frame 792.

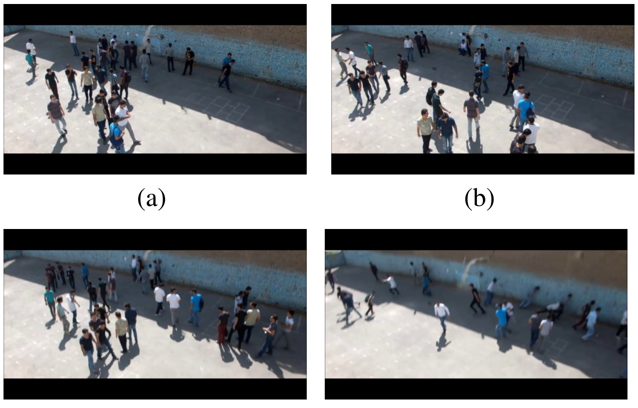
(a)
(b)
(c)
(c)

Figure 5. MED video 7. Normal behavior: (a) frame 100 (b) frame 300. (c) Obstacle avoidance and crowd motion direction change at frame 524. (d) Crowd in panic, very shaky camera at frame 792.

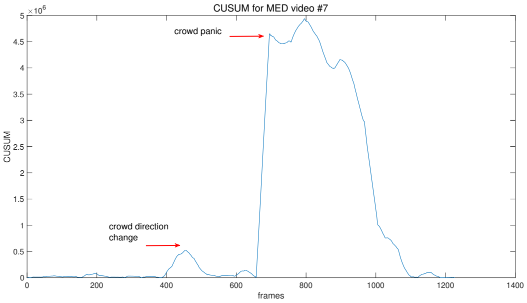

Figure 6. CUSUM values for MED video #7. Peaks correspond to changes in motion direction due to obstacle avoidance and panic.

In our experiments we observed that, in some cases, the proposed method detects changes in the crowd motion and density that are not annotated as ground truth in https://github.com/hosseinm/med/. The videos with no "abnormal events" are a clear case of this, where according to the abnormal even annotations, there is no unusual activity. However, the CUSUM values, depicted in Fig. 8, show changes in the motion distribution. It should be noted that this figure differs from the CUSUM plots corresponding to the synthetically generated videos in Sec. 4.1, as it comprises of real world data with several changes in the motion distribution, varying illumination and density in the scene.

Upon examination of the video itself, we can see that there are indeed variations in the crowd density and motion at changepoints detected by our approach, that are not considered abnormal events in the ground truth. In Fig. 7 we show the frames of video 24 before and at the changepoints detected by our algorithm, to show that there is a difference in motion and density, even though there is not one of the annotated abnormal events. At frame 462 there is no unusual event, however the crowd density has decreased significantly, as most people have walked away from the center of the frame towards its edges. At frame 760 the crowd density in the center of the frame has increased again, and a man suddenly stretches his arms out towards his friend, they approach and greet each other, and form a small cluster of activity. The crowd gradually thins out again near frame 1160.

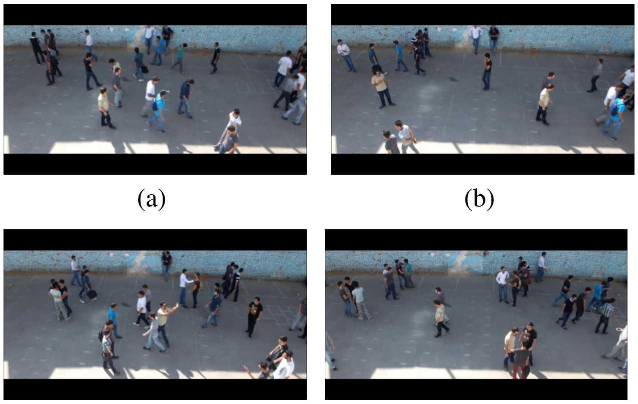
(a)
(b)
(c)
(c)

Figure 7. MED video 24 (a) frame 250: normal motion. No ground truth change, our method finds a change in (b) frame 462: corresponding to decreased crowd density, (c) frame 760: corresponding to sudden arm motion of friends meeting and forming a denser cluster of people, (d) frame 1160: crowd density has decreased as people are moving towards the corners of the video frame.

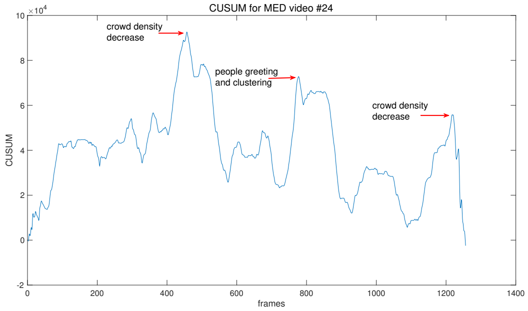

Figure 8. CUSUM values for MED video #24. Peaks correspond to changes in crowd density and sudden motions that have not been annotated in the ground truth as separate events.

5. Conclusions

This work presents a transform-domain approach to abnormal event detection in videos featuring dense motion, which can be approximated by a random distribution. This provides a description of the motion statistics in the video,

Video	Ground Truth [8]	our results
Test1	829, 1515, 1600	830, 1336, 1590
Test2	625, 1170, 1270	624, 1156, 1286
Test3	550, 900, 1010	494, 756, 984
Test4	855, 1040	850, 1032
Test5	810, 960	786, 1056
Test6	650, 942	636, 1028
Test7	518, 845	524 , 792
Test8	690, 898	654, 898
Test9	860, 1005	872, 1010
Test10	1290, 1420, 2095, 2293	1258, 1532, 2094, -
Test11	1464, 1625, 1950, 2152	1292, 1698, 1956, 2074
Test12	199, 500, 615	214, 488, 664
Test13	305, 505	260, 490
Test14	110	104
Test15	373	372
Test16	865	860
Test17	760	634
Test18	475	520
Test19	1230	1232
Test20	1285, 1569, 1675	1286, 1444, 1706
Test21	-	786, 984, 1174
Test22	-	664, 864, 1020
Test23	-	528, 970, 1208
Test24	-	462, 760, 1160
Test25	600, 1077	580, 1148
Test26	865, 1325	858, 1398
Test27	640, 818, 1017, 1150, 1412, 1527	534, 740, 1010, 1230
Test28	795, 855	700, 836
Test29	968, 1024, 1197, 1292	866, 1016, 1170, 1294
Test30	1086	1026
Test31	830	866

Table 1. Comparison with ground truth on the MED datasets

without requiring the estimation of optical flow. It is appropriate for cases where appearance features are not informative, for example in stochastically moving textures with a homogeneous appearance. Sequential change detection is applied to detect changes in such videos, corresponding to changes in motion, and consequently to abnormal events. We perform experiments with homogeneously textured video frames to determine the method's performance for different cases of changes in the stochastic motion of homogeneously textured videos. We also carry out experiments on the MED dataset, introduced in 2016, where our method detects changes in activities with accuracy, while also detecting changes in the crowd motion and density that do not correspond to specific events. This is achieved with no estimation of optical flow values, with no training and no prior knowledge of the activities in the dataset. Future work includes expanding this approach to a wider range of cases of real world crowd or dynamic texture videos, and supplementing it with deep learning layers for the extraction of rich mid-level descriptive features, aiming at its extension to recognition problems. It is expected to yield a lower computational cost through the use of motion statistics, and increase event detection robustness due to its independence from optical flow estimates.

References

[1] H. A.-A. D. Helbing, A. Johannson. Crowd turbulence: the physics of crowd disasters. In *International Conference on Nonlinear Mechanics*, 2007.

[2] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *Int. J. Comput. Vision*, 51(2):91–109, Feb. 2003.

[3] S. Dubois, R. Peteri, and M. Menar. Characterization and recognition of dynamic textures based on the 2d+t curvelet transform. *Signal, Image and Video Processing*, 9(4):819 – 830, 2015.

[4] B. Ghanem and N. Ahuja. Phase based modelling of dynamic textures. In *2007 IEEE 11th International Conf. on Computer Vision*, 2007.

[5] 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, 2015.

[6] E. S. Page. Continuous inspection scheme. *Biometrika*, 41:100–115, 1954.

[7] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 2nd edition, 1987.

[8] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino. Novel dataset for fine-grained abnormal behavior understanding in crowd. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 95–101, 2016.

[9] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe. Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. *CoRR*, abs/1610.00307, 2016.

[10] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe. Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[11] J. Shao, C. C. Loy, K. Kang, and X. Wang. Slicing convolutional neural network for crowd video understanding. In *CVPR*, pages 5620–5628. IEEE Computer Society, 2016.

[12] N. Shroff, P. Turaga, and R. Chellapa. Moving vistas: Exploiting motion for describing scenes. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1911 – 1918. IEEE, IEEE, 2010/06/13/18 2010.

[13] J. Smith, C.-Y. Lin, and M. Naphade. Video texture indexing using spatio-temporal wavelets. In *Proceedings. International Conference on Image Processing*, 2002.

[14] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, pages I–190–I–198, 2014.

[15] R. Vidal and A. Ravichandran. Optical flow estimation and segmentation of multiple moving dynamic textures. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:516–521 vol. 2, 2005.

[16] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31:539?555, 2009.

[17] V. M. Weixin Li and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence (TPAMI), 36(1):18–32, Jan. 2014.