ABNORMAL EVENT DETECTION IN CROWDED SCENES BASED ON STRUCTURAL MULTI-SCALE MOTION INTERRELATED PATTERNS

Dawei Du — Univ. of Elec. Sci. and Tech. of China, daviddo@yahoo.cn

*Honggang Qi, Qingming Huang** — Univ. of the Chinese Acad. of Sci., {hgqi,qmhuang}@ucas.ac.cn

Wei Zeng — NEC Laboratories, China, zeng_wei@nec.cn

Changhua Zhang — Univ. of Elec. Sci. and Tech. of China, zhangch@uestc.edu.cn

ABSTRACT

Detecting abnormal events in crowded scenes remains challenging due to the diversity of events defined by various applications. Among the many application situations, motion analysis for event representation is suited for crowded scenes. In this paper, we propose a novel abnormal event detection method via likelihood estimation of dynamic-texture motion representation, called Structural Multi-scale Motion Interrelated Patterns (SMMIP). SMMIP combines both original motion patterns and their structural spatio-temporal information, which effectively represents localized events by different resolutions of motion patterns. To model normal events, the Gaussian mixture model is trained with the observed normal events, then the likelihood estimation for testing events is computed to judge whether they are abnormal. Meanwhile, the proposed model can be learned online by updating the parameters incrementally. The proposed approach is evaluated on several publicly available datasets and outperforms several other methods proposed before, which is shown that the structural spatio-temporal information added in motion representation helps increasing the anomalies detection rate.

Index Terms— Abnormal Event Detection, Structural Multi-scale Motion Interrelated Patterns, Gaussian Mixture Model

1. INTRODUCTION

Video surveillance plays an extremely important role in the security monitoring of the square, train station, bank and airports, etc. Among the many applications, detecting the abnormal events automatically in the video instead of human labour is considerably important and valuable. Abnormal event detection refers to the problem of finding patterns in video data that do not conform to expected events [1], such as non-expected objects and irregular pedestrians moving style, etc. However, abnormal event samples are usually difficult to collect before, thus our task is a one-class classification problem, or to detect the anomalies based on normal samples.

In crowded scenes, trajectories-based anomaly detection methods [2, 3] may cause miss or false detections because of usual occlusions and inappropriate interactions among multiple objects. Therefore motion-based methods [4, 5, 6, 7, 8, 9, 10] are more generally employed in the research. On the other hand, appearance-based methods are not suited for crowded scenes to distinguishing events due to the changing appearances over time. However existing motion-based methods [7, 8, 9, 10] mostly depend on optical flow, which only captures motions for successive frames and is inaccurate especially in textureless regions [11]. Consequently, better motion representation method should capture motions in different spatio-temporal scales simultaneously even in textureless regions.

In this work, in order to represent different resolutions of motion patterns better, we propose the Structural Multi-scale Motion Interrelated Patterns (SMMIP) which combines motion features and their structural spatio-temporal information. To avoid overfitting problem, too high dimension of original feature is reduced by PCA-whitening method. Then the Gaussian mixture model (GMM) is built to learn normal motion patterns, and the likelihood estimation is computed to judge whether one space-time patch is abnormal. To efficiently adapt new video data streams, we employ an online updates for the GMM model. The proposed approach is evaluated in several publicly available datasets and outperforms several other methods proposed before. This paper has the following two contributions:

- Motion Interchange Patterns (MIP) representation [12] can capture more reliable motion features than optical flow. To the best of our knowledge, this is the first attempt to apply MIP representation to abnormal event detection field;
- The Original MIP representation is combined with structural spatio-temporal information of motion patterns, namely SMMIP, which helps the learning model capture different resolutions of motion patterns better.

*This work was partially funded by National Natural Science Foundation of China: 61025011 and 61001108, partially funded by National Basic Research Program of China (973 Program): 2009CB320906.

The influence on the detection rate by different parameters is analyzed in detail and the online GMM is updating the parameters incrementally to improve the results.

The rest of this paper is organized as follows. Section 2 discusses related works in this field, Section 3 describes the proposed method for abnormal event detection, Section 4 presents experimental results on several publicly available datasets, and Section 5 presents conclusion for our work.

2. RELATED WORKS

Current algorithms for abnormal event detection employ three low-level representation schemes, namely the local descriptor (STIP, etc.), optical flow, and dynamic-texture (LBP-TOP, MDT, etc.). Zhao *et al.* [13] seek coordinates of space-time interest points (STIP) and then extract HOG/HOF descriptors around each such point. STIP has proven effective in action recognition or event classification field, yet the chief drawback is that the technique relies on the assumption that one can reliably detect a suitable number of stable interest points: too few may not provide enough information, while too much may drown any informative cues. Cong *et al.* [7] estimate multi-scale HOF in space-time patches to represent the events. Optical flow provides strong cues for capturing the local dynamics, but it commits early-on to a particular motion estimate at each pixel, and unreliable or wrong estimates would provide misleading information. Zhao and Pietikainen [14] extend the local binary patterns descriptor to 3D video time-varying dynamic textures, called Local Binary Patterns-Three Orthogonal Planes (LBP-TOP). And it is successfully applied to abnormal event detection [15]. Mahadevan *et al.* [6] model appearance and dynamics in the video scene by the mixture of dynamic texture (MDT) [16].

Meanwhile, A variety of learning models for low-level representation have been proposed. Clustering approaches [2, 17, 18] have been widely used to detect unusual events because of their good performance, where the detection is carried out by finding spatially isolated points. Adam *et al.* [10] monitor optical flow directions and speeds at fixed spatial locations, and determine the likelihood of new observations. Mehran *et al.* [8] employ the social force model to formulate the abnormal crowd behavior and then use Latent Dirichlet Allocation to detect anomalies. Kim and Grauman [9] model local optical flow with MPPCA and enforce the consistency by MRF. In [5], the spatio-temporal features are modeled as Gaussian distributions. And a set of prototypes are maintained for each location in a scene, which serve as states for a coupled hidden Markov model. In [19], a graph-based nonlinear dimensionality reduction method is used for abnormality detection. Cong *et al.* [7] present an algorithm for anomaly detection based on the sparse reconstruction cost.

As said before, MIP representation can capture more reliable motion features than optical flow representation, also has been applied in action recognition successfully [12]. Further, in order to model different resolutions of motion patterns in small space-time region better, SMMIP representation combines original MIP features with their structural spatio-temporal information.

3. ABNORMAL EVENT DETECTION BY SMMIP

The objective is to infer when and where abnormal events happen in crowded scenes. Firstly, SMMIP feature is extracted from space-time patches. Secondly, too high dimension of original feature space is reduced by PCA-whitening method to avoid overfitting. Thirdly, GMM is employed to learn the normal motion patterns and detect the anomalies by online updating the parameters incrementally.

3.1. Structural Multi-scale Motion Interrelated Patterns

For an input space-time patch, its motion feature is represented by 8-trinary-digit string of 8 channels in every pixel. Each digit compares the compatibility of two motions with the local patch similarity pattern: one motion in a specific direction from the previous frame to the current frame, and another in a different direction from the current frame to the next one. A value of +1 indicates that the former motion is more likely, -1 indicates that the latter is more likely. A values of 0 indicates that both are compatible in approximately the same degree. The design of three patches enables the comparison of two similarly scaled data points to capture the motion of the small region around the point. The encoding scheme based on three patches is depicted in Fig.1.

As seen in Fig.1, encoding is performed for every pixel in every triplet of frames (previous, current, and next frame). The SSD templates are picked from varying $s_n \times s_n$ pixel patches in spatial scale, and their SSD scores for comparing the compatibility of two motions are computed as follows:

$$\begin{cases} D_1 = \sum_{m=1}^{s_n} \sum_{n=1}^{s_n} [I(m,n,t_p(1)) - I(m,n,t_p(2))]^2 \\ D_2 = \sum_{m=1}^{s_n} \sum_{n=1}^{s_n} [I(m,n,t_p(3)) - I(m,n,t_p(2))]^2 \end{cases}. \quad (1)$$

where temporal scale $t_p = [-p, 0, p]$ is the temporal location for the triplet of frames, if the location of current frame is denoted as the origin. D_1 and D_2 means the SSD scores, namely the sum of squared differences between the patch in the previous and next frame gray image $I(m, n, t)$, respectively. Thus different combinations of $\{s_n, t_p\}$ represent different resolutions of motion patterns in spatio-temporal scales.

All combinations of i and j are used for encoding and the angle α between i and j is $0°, 45°, \cdots, 315°$. For each direction channel α, the motion pattern is therefore represented by a 64-trinary-digit code $S_{i,j}(\alpha)$, which is computed as follows:

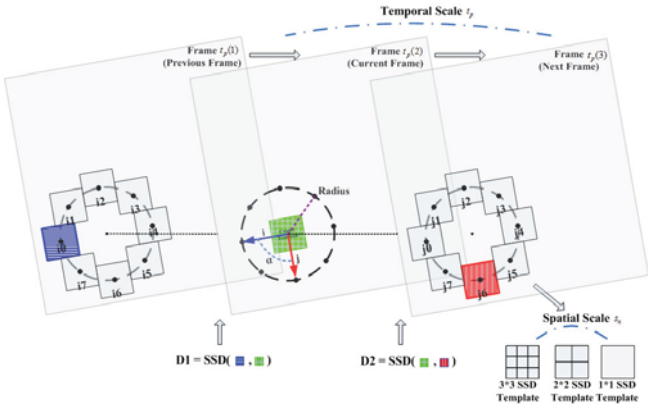


Fig. 1. The basic encoding level. Relative to the location of the green patch with grid (center template) in the current frame, the location of the patch in the previous and the next frame is said to be in direction i and j respectively. The different motion patterns captured by different directions are identified from 0 to 7, such as the blue patch with horizontal line $i0$ in the previous frame and the red patch with vertical line $j6$ in the next frame. Better view is obtained in color version.

$$S_{i,j}(\alpha) = \begin{cases} +1 & if \quad D_1 - D_2 > \Theta \\ 0 & if \quad |D_1 - D_2| \leq \Theta \\ -1 & if \quad D_1 - D_2 < -\Theta \end{cases} . \quad (2)$$

Following the suppression mechanism [12], we are left with 8 channel maps, where each pixel is encoded by one 8-trinary-digit string per channel. We treat the positive and negative parts of the string separately and obtain a 16-bit one $C(\alpha) = [C_{pos}(\alpha), C_{neg}(\alpha)]$. Thus these 16 values represent the complete motion patterns for that pixel. For each channel, the frequency for every kind of code in C_{pos} and C_{neg} is collected in small patches respectively to create 512-dimensional frequency histogram.

As seen in our experiments, despite the success of MIP in encoding motion, its limited scale characterization is outperformed by the multi-scale feature representation employed by our method. Thus the final motion feature vector V is

$$V = [F_{s_1,t_1}, \cdots, F_{s_m,t_n}] \quad if \quad \sum F_{s_i,t_j} > T_F, \quad (3)$$

where F_{s_i,t_j} is the 512-dimensional frequency histogram in spatial scale $s_i (i = 1, \cdots, m)$ and temporal scale $t_j (j = 1, \cdots, n)$. Because the anomalies can only happen in motion region, the threshold value T_F is utilized to extract the efficient foreground motion, which can reduce noise and improve the efficiency.

At the basic encoding level, our method and MIP [12] both share the use of three patches and two SSDs. However, there are several differences between the two methods:

- MIP is restricted to the fixed spatial and temporal scale of SSD templates and triplet of frames, respectively. Therefore MIP is not so general to capture the different resolutions of motion patterns well simultaneously. However, SMMIP is more general to represent them by combining the original fixed-scale MIP representation and its several structural spatio-temporal scales information.

- After basic encoding level, MIP employs k-means and bag-of-words approach to represent each video clip for action recognition by eight histograms (one per channel), while we focus on the representation of each space-time patch for abnormal event detection by concatenating all the eight frequency histograms in every spatio-temporal scale we consider, which is more discriminating for representing motions of small patch than bag-of-words method in high dimension.

In Eq.(3), SMMIP code considers multiple spatial or temporal scales when m or n is assigned as plural numbers, and SMMIP code will degenerate to MIP code when $m = 3, n = 1$. Moreover, the radius of the template circle (the distance between the center point of center template and the one of its surrounding templates, viewed as purple dotted line in Fig.1) can be changed if necessary. By these changes our motion representation can be flexibly applied in various situations.

3.2. Preprocessing of High Dimensional Feature

The dimension of the feature vector V is $D_{old} = 8 \times 512 \times m \times n = 4096mn$, which means that even the minimal dimension is 4,096. It is too high for Gaussian model to learn and results in overfitting. In order to avoid the dimensionality of curse we reduce the old high dimension D_{old} to the new one D_{new} by PCA method. Besides, whitening method makes the feature data independent from each other by transforming the covariance matrix to an identity matrix. Thus the Euclidean distance is equivalent to the Mahalanobis distance, which is a more accurate measure of the distance between high dimensional vectors.

3.3. Abnormal Event Detection by Online GMM

In the training period, we employ GMM to model normal motion patterns and estimate its parameters by Expectation Maximization algorithm [20]. To improve the efficiency, k-means method is applied to initialize the clustering centers firstly. The probability density function is

$$L_p = \gamma \cdot log(p(V|\theta)) = \gamma \cdot log(\sum_{i=1}^{K} \pi_i N(V|\mu_i, \Sigma_i)), \quad (4)$$

where $\theta = \{\pi_i, \mu_i, \Sigma_i\}$ denote the weight, mean and covariance of component respectively among all the K components,

and γ is the quantification coefficient. The patch with the likelihood value L_p smaller than the threshold T_p is classified as abnormal. In some specific applications, we can employ the integrating multiple monitors scheme [10] to decrease some unnecessary false alarms.

In order to avoid any issues with concept drift, for every incoming V^t, all the Gaussian components are checked to choose the kth Gaussian with the highest probability in $k = \arg\max_i log(p(V|\mu_i, \Sigma_i)$. After that, we can update the parameters of kth matched Gaussian as follows:

$$\begin{cases} \pi_k^t &= (\pi_k^{t-1} + \Delta\pi)/(\sum \pi_k^{t-1} + \Delta\pi) \\ \mu_k^t &= (1-\beta)\cdot\mu_k^{t-1} + \beta\cdot V^t \\ \Sigma_k^t &= (1-\beta)\cdot\Sigma_k^{t-1} + \beta\cdot(\mu_k^{t-1} - V^t)^T(\mu_k^{t-1} - V^t) \end{cases} \tag{5}$$

where β is the learning rate and $\Delta\pi$ $(0 < \Delta\pi < 1)$ is the reward for this matched kth Gaussian. The rest unmatched Gaussian only update their weights according to $\pi_j^t = \pi_j^{t-1}/(\sum_i \pi_j^{t-1} + \Delta\pi), j \neq k$.

4. EXPERIMENTAL RESULTS

To test the effectiveness of the proposed method, we evaluate it on three publicly available datasets, including the UCSD Ped1 Dataset [6], the UMN Dataset [8] and the Subway Exit Dataset [10]. Our approach has been compared with some recent state-of-the-art methods in terms of equal error rate (EER), rate of detection (RD) and the area under ROC (AUC) in the UCSD Ped1 and UMN Datasets. Meanwhile, an comparison inside our method is reported in Subway Exit Dataset.

4.1. Parameters Discussion

Because the motion feature is trained using several parameter and constants, some choices have to be made in the algorithm described above, and the best results are reported below.

The reduced dimension D_{new} varies from 512 to 1024. The more spatial and temporal scales we utilize, the more reduced dimension we need. So is the number of GMM components K. It is evaluated from 20 to 35 among all the experiments.

SMMIP representations by several spatio-temporal scale combinations $\{s_m, t_n\}$ are designed in the experiment, named as MIP ($\{s_3, t_1\}$), MIP-Time ($\{s_3, t_{1,2}\}$), MIP-Sapce ($\{s_3, t_{1,2}\}$) and MIP-TS ($\{s_{2,3}, t_{1,2}\}$). $s_{2,3}$ means 2×2 and 3×3 SSD template in spatial scale, and $t_{1,2}$ means $[-1, 0, 1], [-2, 0, 2]$ in temporal scale for the triplet of frames.

The threshold Θ for the SSD template is 1296 for s_3 and 576 for s_2 respectively. And the threshold $T_F = 1.6 \cdot p_{height} \cdot p_{width}$, where $\{p_{height}, p_{width}\}$ is the size of every space-time patch.

4.2. UCSD Ped1 Dataset

The UCSD Ped1 Dataset includes some pedestrian scenes on campus, with 34 training clips and 36 test clips. There is a subset of 10 clips in testing set provided with pixel-level binary masks, which identify the regions containing abnormal events. Each clip has 200 frames with a 158×238 resolution. We split each frame into 25×25 local patches without overlapping.

Table 1. Results on UCSD Ped1 dataset

Methods	EER	RD	AUC
SF [8]	79%	21%	17.9%
MPPCA [9]	82%	18%	20.5%
MDT [6]	55%	45%	44.1%
Adam [10]	76%	24%	13.3%
Sparse [7]	54%	46%	46.1%
MIP	50.3%	49.7%	50.0%
MIP-Time	44.5%	55.5%	56.8%
MIP-Space	44.4%	55.6%	57.4%
MIP-TS	**41.3%**	**58.7%**	**64.9%**

Table 1 shows the quantitative comparison of our method with [6, 7, 8, 9, 10], from which we can find that the performance of our methods outperforms the state-of-the-art methods in pixel-level measurement defined in [6]. And RD increases by al least 5% when considering multiple spatial and temporal scales. Although SMMIP representations perform almost the same in frame-level measurement (EER is around 32%, AUC is around 74%), the motion features with their structural spatio-temporal information produces a significant improvement in pixel-level measurement.

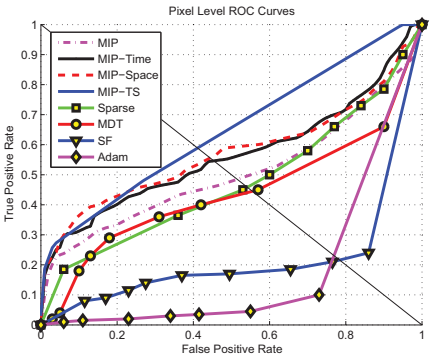


Fig. 2. Pixel-level ROCs of UCSD Ped1 Dataset

In Fig.2, we compare SMMIP representations with other methods such as MDT [6], Sparse [7], SF [8] and Adam [10], in pixel-level measurement. It clearly shows the advantage for considering structural spatio-temporal information. Several detected anomalous objects in test clips are highlighted in red rectangle in Fig.3.

bicycle wheelchair vehicle skater

Fig. 3. Some detected objects in UCSD Ped1 Dataset

4.3. UMN Dataset

The UMN Dataset consists of 3 different scenes of crowded rapid escape events, with 1,450, 4,415 and 2,145 frames for Scene 1, 2, 3, respectively. At first, each image with a 320×240 resolution is converted into gray image and splited into 25×30 sub-regions, then the motion feature is extracted from each sub-region. The GMM is trained from the first 400 regular event frames of each scene with the others for test.

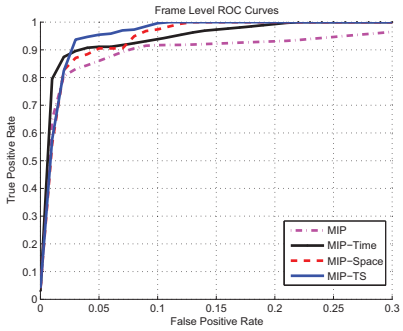


Fig. 4. Frame-level ROCs of UMN Dataset

In Fig.4, the ROC curves of SMMIP representations in frame-level are shown, from which we can find motion representations in multiple scales perform better than the one in single scale. Table 2 provides the quantitative comparisons to the state-of-the-art methods [7, 8, 21]. The AUC of feature in multiple scale is average 98.2% among all the three scenes, while the one in fixed scale is average 94.9%, whose result underscores our contribution in employing structural spatio-temporal information to capture different resolutions of motion patterns. Our approach outperforms [7, 8] and is comparable to [21], which demonstrates the effectiveness of our motion representation method. In Fig.5, some snapshots in 3 scenes are shown by the escape people marked as red regions.

Scene 1 Scene 2 Scene 3

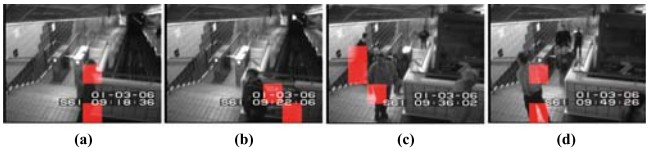
Fig. 5. Sample escape snapshots in UMN Dataset

Table 2. Results on 3 scenes of UMN Dataset

Methods	AUC		
Chaotic Invariants [21]	99%		
SF [8]	96%		
Optical flow [8]	84%		
Nearest Neighbour [7]	93%		
Sparse [7]	99.5%	97.5%	96.4%
MIP	**99.6%**	94.4%	90.8%
MIP-Time	99.3%	96.5%	97.1%
MIP-Space	99.5%	97.2%	96.0%
MIP-TS	99.2%	**97.8%**	**97.8%**

4.4. Subway Exit Dataset

The Subway Exit Dataset is a subway surveillance video in exit gate (43 minutes long with 64,900 frames). For simplicity, the original frames is resized from 512×384 to 320×240 and the new frames are divided into 39×39 local patches without overlapping. Then the first 8 minutes (the first whole round for get-off of the train) are collected to learn the normal motion patterns.



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

Fig. 6. Some detected events in Subway Exit Dataset are shown, where (a) is cleaning, (b) is motion blur, (c) is wrong direction and (d) is watching around.

Table 3. Results on Subway Exit Dataset

Methods	Wrong Direction	Cleaning	Misc	False Alarm
Ground Truth	9	4	3	0
MIP	8	3	2	5
MIP-Time	9	3	2	3
MIP-Space	9	3	3	3
MIP-TS	9	4	3	2

Due to no ground truth labels in this video, other researchers define different types of events respectively [9, 10, 13], therefore it is difficult to compare our method with others. We define 3 different types of abnormal events: 1) walking in the wrong direction; 2) cleaning; 3) misc like watching around and motion blur, etc. There is only the comparison of SMMIP representation below. Quantitative comparison results are shown in Table 3, and some detected abnormal events are shown in Fig.6. Although all the motion representations

can detect the abnormal events we define, motion representation with structural spatio-temporal information results in more accurate detected regions and less false alarms.

5. CONCLUSION

This paper proposed a novel Structural Multi-scale Motion Interrelated Patterns for abnormal event detection based on the Gaussian likelihood estimation. By incrementally updating the components of the Gaussian mixture model, our method supports online event detection. In order to capture different resolutions of motion patterns, the motion representation is combined with its structural spatio-temporal information. Experimental results on several publicly available datasets demonstrate the effectiveness of the proposed algorithm, and the performance can be improved significantly compared to the state-of-the-art methods. We believe that this kind of motion patterns in this work can also apply to other applications, such as event or action recognition.

6. REFERENCES

[1] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *IJCV*, vol. 74, no. 1, pp. 17–31, Aug. 2007.

[2] O. Arandjelovic, "Contextually learnt detection of unusual motion-based behaviour in crowded public spaces," *ISCIS*, pp. 403–410, 2011.

[3] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *TCSVT*, vol. 18, no. 11, pp. 1544–1554, 2008.

[4] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, "A system for learning statistical motion patterns," *TPAMI*, vol. 28, no. 9, pp. 1450–1464, 2006.

[5] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," *CVPR*, pp. 1446–1453, Jun. 2009.

[6] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," *CVPR*, pp. 1975–1981, Aug. 2010.

[7] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," *CVPR*, pp. 3449–3456, Jun. 2011.

[8] R. Mehram, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *CVPR*, pp. 935–942, Jun. 2009.

[9] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," *CVPR*, pp. 2921–2928, 2009.

[10] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *TPAMI*, vol. 30, pp. 555–560, Mar. 2008.

[11] R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer, 2011.

[12] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," *ECCV*, pp. 256–269, 2012.

[13] B. Zhao, L. Fei-Fei, and E. P. Xing, "Oline detection of unusual events in videos via dynamic sparse coding," *CVPR*, pp. 3313–3320, 2011.

[14] G. Zhao and M. Pietikainen, "Dynamic texture recognition to facial expressions," *TPAMI*, vol. 29, no. 6, pp. 915–928, Aug. 2007.

[15] J. Xu, S. Denman, S. Sridharan, C. Fookes, and R. Rana, "Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes," *Joint ACM Workshop on Modeling and Representing Events*, 2011.

[16] A. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *TPAMI*, vol. 30, no. 5, pp. 909–926, May. 2008.

[17] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," *CVPR*, 2004.

[18] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. lsbell, and G. Goleman, "Detection and explanation of anomalous activeities: Representing activities as bags of event n-grams," *CVPR*, 2005.

[19] I. Tziakos, A. Cavallaro, and L. Xu, "Event monitoring via local motion abnormality detection in non-linear subspace," *Neurocomputing*, 2010.

[20] C. M. Bishop, *Pattern Classification and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., 2006.

[21] S. Wu, B. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," *CVPR*, pp. 2054–2060, 2010.