2014 22nd International Conference on Pattern Recognition

Anomaly Detection through Spatio-Temporal Context Modeling in Crowded Scenes

Tong Lu, Liang Wu, Xiaolin Ma
National Key Laboratory of Novel Software Technology
Nanjing University, Nanjing, China 210023
Email: lutong@nju.edu.cn

Palaiahnakote Shivakumara
Faculty of Computer Science and Information Technology
University of Malaya, Kuala Lumpur, Malaysia
Email: hudempsk@yahoo.com

Chew Lim Tan
School of Computing, National University of Singapore
Email: tancl@comp.nus.edu.sg

***Abstract*—A novel statistical framework for modeling the intrinsic structure of crowded scenes and detecting abnormal activities is presented in this paper. The proposed framework essentially turns the anomaly detection process into two parts, namely, *motion pattern representation* and *crowded context modeling*. During the first stage, we averagely divide the spatio-temporal volume into atomic blocks. Considering the fact that mutual interference of several human body parts potentially happen in the same block, we propose an atomic motion pattern representation using the Gaussian Mixture Model (GMM) to distinguish the motions inside each block in a refined way. Usual motion patterns can thus be defined as a certain type of steady motion activities appearing at specific scene positions. During the second stage, we further use the Markov Random Field (MRF) model to characterize the joint label distributions over all the adjacent local motion patterns inside the same crowded scene, aiming at modeling the severely occluded situations in a crowded scene accurately. By combining the determinations from the two stages, a weighted scheme is proposed to automatically detect anomaly events from crowded scenes. The experimental results on several different outdoor and indoor crowded scenes illustrate the effectiveness of the proposed algorithm.**

I. Introduction

Nowadays, automatic analysis of densely crowded environments such as subways, universities, railway stations and stadiums has been a recent interest in pattern recognition. Methods for crowed modeling [1], crowed size estimation [2] and individual behavior prediction [3] have been proposed. Most crowed monitoring systems aim at assisting public security, thus these efforts face a common problem of detecting deviations from crowded environments. The problem is referred to as crowded anomaly motion detection, which provides much more valuable hints than detecting normal behaviors. The anomaly motions here are defined as the unusual events that are different from those steady ones frequently happen at particular positions inside a scene.

Unlike anomaly detection from non-crowded scenes, a crowed environment generally requires monitoring an excessive number of individuals and their activities through video surveillance devices. As a result, computational approaches of anomaly detection in densely crowded scenes may face more difficulties both in scene modeling and anomaly behaviors detecting in two aspects. First, multiple individuals in a crowded scene in general lead to severe occlusions, which make object tracking fairly difficult and sometimes a significant challenge even for human observers. Second, highly irregular pedestrian behaviors within the same crowded scene potentially make explicit modeling of anomaly deviations difficult, and therefore it may be hard to distinguish tiny deviations from normal behaviors clearly.

In this paper, we present a novel spatio-temporal framework for modeling the intrinsic structure of crowded scenes and detecting abnormal activities in it by using Gaussian Mixture Model (GMM) and Markov Random Field (MRF). Since each crowded scene essentially consists of a large number of pedestrian activities together with their interrelations, we turn the anomaly detection process into two parts, namely, *motion pattern representation* and *crowded context modeling*. During the first stage, we averagely divide the spatio-temporal volume of the scene into atomic blocks. Considering the fact that the mutual interferences of several human body parts potentially happen in the same block, we propose an atomic motion pattern representation using GMM to distinguish the motions inside each block in a refined way. Usual motion patterns can thus be defined as a certain type of steady motion activities appearing at specific scene positions. During the second stage, we further use the MRF model to characterize the joint label distributions over all the adjacent local motion patterns inside the same crowded scene, aiming at modeling the severely occluded situations in the crowded scene accurately. By combining the determinations from the two stages, a weighted scheme is finally proposed to automatically detect anomaly events in crowded scenes.

The main contribution of our approach is to present a statistical approach by coupling spatio-temporal crowded context modeling for anomaly event detection in crowded scenes. The context model is suitable to characterize the relations between spatially or temporally adjacent pedestrian activities, thus enforcing local consistency on them to assist detecting anomaly motion patterns. The experimental results on several real-life outdoor and indoor scenes illustrate the effectiveness of the proposed method by outperforming the existing anomaly detection algorithm.

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.383

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 gives the representation of atomic motion patterns of crowded scene using GMM, then Section 4 models the context for the activities in crowded scene by MRF. Experiments and discussions are given in Section 5. Finally, Section 6 concludes the paper.

II. Related Work

A number of methods have been proposed in the past years and the existing methods for anomaly detection can be roughly classified into three categories: *trajectory learning*, *motion modeling*, and *unsupervised approach*.

Generally, a *trajectory learning approach* first tracks each object in a scene and then learns a model from object tracks to identify unusual activities by detecting deviations [4]. Every scene object is tracked over a dynamic sequence of inferred tracking states as $S_T = \{s_1, s_2, ...s_T\}$, where T is a frame sequence and s_t depicts things such as appearance, position, shape, velocity and direction. In [5], an anomaly detection system is proposed to automatically learn motion patterns by tracking multiple objects, in which growing and prediction of cluster centroids of foreground pixels ensure the accuracy of tracking a moving object in the scene. Unfortunately, it is not very promising for tracking objects in densely crowded scenes. The method in [6] is one of the first algorithms for tracking objects in a crowded environment, which uses articulated ellipsoids to model human appearance and a Gaussian distribution to model the background for segmentation. Recently, Ali and Shah [3] present a force model which has three floor fields inspired by the research in the field of evacuation dynamics, namely Static Floor Field (SFF), Dynamic Floor Field (DFF) and Boundary Floor Field (BFF). However, it is still challenging to tracking individuals in a crowed scene by simply using trajectory learning techniques.

Motion modeling is an alternative approach which has been proposed by modeling motion patterns to avoid object tracking [7]. This approach focuses on distinguishing unusual events from other stationary behaviors and has been proved suitable for anomaly detection especially in crowded scenes. In this approach, optical flow is a popular low-level representation to describe motion patterns from surveillance video. Andrade *et al.* [8] implement spectral clustering on feature prototypes that are obtained by performing PCA on the optical flow fields, and train the MOHMM model for each class during anomaly detection. Adm *et al.* [9] calculate the probabilities of optical flows in local regions. Unfortunately, optical flow, together with other similar low-level descriptors such as pixel change histograms and background subtraction operations, are not reliable enough for detecting abnormal events from crowded scenes in which occlusions always exist [10].

Since no prior assumptions of what unusual events may look like are required, *unsupervised anomaly detection* can be considered as a complementary but simultaneously a more natural approach to detect unusual events. It is especially suitable for the situations such as the lack of sufficient training data, and the volatility of defining normality and abnormality. Xiang and Gong [11] propose natural grouping on behavior patterns through an unsupervised model, which selects eigenvector features from a normalized affinity matrix. Zhao *et al.* [12] propose a fully unsupervised dynamic sparse coding approach for detecting unusual events in videos. Most of the unsupervised methods utilize the "*hard to describe*" but "*easy to verify*" property of unusual events, without building an explicit model for normal events. Its advantage is that one can compare each event with all the other observed events and thus determine whether a given event is abnormal. However, they still have the same problems, namely, how to improve the accuracy and how to efficiently measure the similarity of the detected events, especially on a relatively large dataset that is required by most unsupervised methods.

III. Representation of Atomic Motion Patterns using GMM

Due to the relatively large number of individuals in a crowded scene, segmentation according to the contours of pedestrians in the scene is not always reliable as discussed. Averagely dividing the spatio-temporal volume into atomic spatio-temporal blocks with a fixed size is popular to model motions in a local spatio-temporal block. However, it is still difficult to avoid the discordance of motions within the same block. Namely, even for the same block, it potentially contains discordance of motions which are generally brought by different body parts of the same pedestrian, or the body parts from different pedestrians that are unexpectedly divided into the same block. As a result, using a uniform representation to characterize such a single spatio-temporal block in general faces difficulties in characterizing multiple motions inside the same block, especially considering the mutual interferences of body parts from the same pedestrian or several different neighboring pedestrians.

In order to better characterize each spatio-temporal volume block, we define the motion pattern within each block as a collection of motions from multiple body parts. For this purpose, we first extract optical flow vectors and then employ the Gaussian Mixture Model over the optical flow vectors in each block. As a result, every motion component within a spatio-temporal block can be captured and the interference between each other are accordingly avoided since the distribution of optical flows located in that block can be modeled by GMM. The Gaussian Mixture models of all the optical flows within the same block are accordingly named as its motion patterns. In this way, we obtain the atomic motion representation for every block in the whole spatio-temporal volume.

Specifically, we first divide the volume into spatio-temporal blocks of a fixed size as in [13]. We set the size of each block to the average width of pedestrians. For every feature point (x, y) in an input video frame, the optical flow v is computed using the pyramidal Lucas-Kanade algorithm:

$$v = (v_x, v_y) \tag{1}$$

where v_x and v_y are the velocities along the x and y directions, respectively. The flow vectors that have a too large magnitude are considered as noises and thereby discarded directly. Following the described spatio-temporal volume representation, all the optical flow vectors are then assigned into spatio-temporal blocks according to their spatio-temporal locations.

Then for a block at position (i, j) and time moment t, we calculate the Gaussian Mixture Model $G^t_{ij}(x^t_{ij}|\theta^t_{ij})$ for the

block by:

$$G_{ij}^t(x_{ij}^t|\mu_{ij}^t,\theta_{ij}^t)=\sum_{k=1}^{k_{ij}^t}\omega_k g_k(x_{ij}^t|\mu_{ij}^t,\Sigma_{ij}^t) \tag{2}$$

where x_{ij}^t is a 2-dimensional optical flow vector in the block, ω_k $(k=1,...,k_{ij})^t$ denotes a corresponding mixture weight, and $g(x_{ij}^t|\mu_{ij}^t,\Sigma_{ij}^t), k=1,...,k_{ij}^t$ are the component Gaussian densities.

There are altogether k_{ij}^t components in G_{ij}^t, each denoting the motion of a particular body part from the same pedestrian or several different pedestrians. To determine the value of k_{ij}^t, we apply a mean shift clustering algorithm over the optical flows which are assigned to the block at (i,j), and accordingly use the number of the result clusters to initialize k_{ij}^t. Then, given the training optical flow vectors in any spatio-temporal block, we model the motion pattern for it by GMM, namely, the parameter of θ which best matches the distribution of the training optical flow vectors. We use the Maximum Likelihood (ML) algorithm to estimate the parameter. Since the likelihood $P(x|\theta)$ is a non-linear function for the parameter θ and a direct maximization will be difficult to calculate, we use the Expectation-maximization (EM) algorithm to find the maximum likelihood solution. Suppose the model parameters of ω,μ and Σ are denoted by θ:

$$P(x|\theta)=P(x|\omega,\mu,\Sigma)=\prod_{n=1}^{N}\sum_{k=1}^{k_{ij}^t}\omega_k\mathcal{N}(x_n|\mu_k,\Sigma_k) \tag{3}$$

The clustering centers and the number of clusters k_{ij}^t described above are passed to the training process of GMM as the input for initialization. On each EM iteration (M-step), the following re-estimation formulas are used, which guarantee a monotonic increase of the likelihood value in the model:

$$\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n \tag{4}$$

$$\Sigma_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T \tag{5}$$

$$\omega_k=\frac{N_k}{N} \tag{6}$$

where

$$N_k=\sum_{n=1}^{N}\gamma(z_{nk}) \tag{7}$$

$$\gamma(z_{nk})=\frac{\omega_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\omega_j\mathcal{N}(x_n|\mu_j,\Sigma_j)} \tag{8}$$

μ_k,Σ_k and ω_k are mean, covariances and mixing coefficients (of ...), respectively. x_n is an optical flow vector belonging to the spatio-temporal block. In this way, we obtain the model parameters θ if the convergence criterion is satisfied.

We further name the cluster of the motion patterns inside the same block as a *motion prototype*, which is considered as an usual event prototype in the block. Since the mean of a set of GMMs do not always remain a GMM and directly computing the mean of GMMs will be infeasible, we define the medoid M_s^{ijk} of a motion prototype by

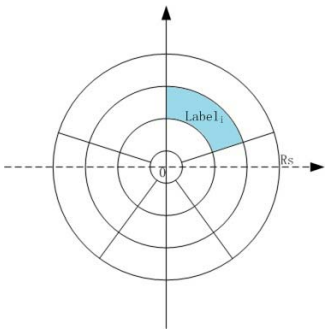


Fig. 1. The spherical polar coordinate space for labeling motion vectors, where the optical flow in the colorful region is assigned a label l_i.

$$M_s^{ijk}=\arg\min_{GMM_k^{ijk}}\sum_{m\neq k}KL(GMM_k^{ijk}||GMM_m^{ijk}) \tag{9}$$

where GMM_k^{ijk} is the motion pattern that is associated with the prototype M_s^{ijk} at position (i,j,k). Since each prototype essentially represents a certain type of steady motion activities which appear in a specific local region inside the scene, we treat it as the *usual* motion pattern at that position. Then we determine whether a given motion pattern is usual or unusual by measuring its deviations from the calculated prototypes through the following confidence:

$$C_{basic}=\max_s(p(G_t^{ijk}|P_s^{ijk})) \tag{10}$$

IV. Context Modeling for Crowded Scenes

Our next purpose is to improve the accuracy of crowded anomaly detection by characterizing the context of motion patterns. Inspired by the work of Galleguillos et. al [14] for object categorization, we use the MRF model to learn the joint label distributions of adjacent local motion patterns as the context of a crowded scene. Each component C_i in GMM is actually a kind of composition of local motion pattern representation and is defined by a Gaussian distribution together with a weighted value ω_i. Thus we can assign a discrete motion label by transforming the mean vector of C_i from the original video spatio-temporal coordinate space to the planar polar coordinate space, which is divided into a set of discrete radius and radian intervals with a fixed size as shown in Fig. 1. Namely, each cluster is assigned with a label of a specific bin in which the centroid vector falls in. To decide the proper size for the intervals, we choose the maximum speed value R_s of pedestrians in the scene s, and correspondingly divide the planar polar coordinate space (r,θ), $r\in(0,R_s)$, $\theta\in(0,360)$ into fixed radius and angle intervals.

For a specific motion pattern that is composed of k_{ij}^t components, we obtain a discrete distribution of labels as shown in Table I.

Next, to describe the spatio-temporal context of motion patterns, we define a set of relation matrices

TABLE I. THE LABEL DISTRIBUTION OF A MOTION PATTERN. $l_i(1 \le i \le k_{ij}^t)$ IS THE LABELS FOR CLUSTERS OF THE MOTION PATTERN, $p_i(1 \le i \le k_{ij}^t)$ IS THE CORRESPONDING PROBABILITY, AND k_{ij}^t IS THE TOTAL NUMBER RELATED TO THE MOTION PATTERN (GMM).

l_1	l_2	l_3	...	$l_{k_{ij}^t}$
p_1	p_2	p_3	...	$p_{k_{ij}^t}$

$\Phi_x^{mn}, \Phi_y^{mn}, \Phi_z^{mn}, \Phi_t^{mn}$, each capturing the interactions between the label distributions from adjacent blocks along one of the 4 directions. Taking the x direction as an example, each entry (i, j) in the matrix Φ_x^{mn} actually represents the *virtual* times of the motion pattern having the label of l_i, which appears in the training volume together with another motion pattern having the label of l_j. The term *virtual* here is a float value rather an integer one to calculate co-occurrence times. To simplify the computation of virtual times between two adjacent motion patterns, we assume that their label distributions are independent. Thus, the joint distribution of labels of any two adjacent motion patterns can be calculated as shown in Table II:

TABLE II. THE JOINT LABEL DISTRIBUTION OF TWO ADJACENT MOTION PATTERNS m AND n. p_i^m IS THE PROBABILITY OF LABEL l_i IN THE LABEL DISTRIBUTION L_m, AND $k_{ij_m}^t$ IS THE TOTAL NUMBER OF LABELS WITHIN THE MOTION PATTERN m.

$p_1^m p_1^n$	$p_1^m p_2^n$	...	$p_1^m p_{k_{ij_n}^t}^n$
$p_2^m p_1^n$	$p_2^m p_2^n$	...	$p_2^m p_{k_{ij_n}^t}^n$
...	...	...	...
$p_{\lvert\mathcal{L}_m\rvert}^m p_1^n$	$p_{k_{ij_m}^t}^m p_2^n$	...	$p_{k_{ij_m}^t}^m p_{k_{ij_n}^t}^n$

After obtaining the probabilities which are extracted from each specific volume, we can empirically count the virtual co-occurrence times between any two adjacent motion patterns. Basically, for one label pair (l_i, l_j), supposing the probability of l_i in the label distribution L_1 is $p^1(l_i)$ and the probability of l_j in the label distribution L_2 is $p^2(l_j)$, we add $p^1(l_i)p^2(l_j)$ to the entry (i, j) in the relation matrix, which corresponds to the two adjacent motion patterns. Therefore, the entry (i, j) in matrix Φ_x^{mn} can be calculated as following:

$$\Phi_x^{mn}(i,j) = \sum_{k=1}^{|\mathcal{D}|} p_k^m(l_i) p_k^n(l_j) \tag{11}$$

where $|\mathcal{D}|$ is the total number of the volumes obtained from the training image sequence, $p_k^m(l_i)$ stands for the probability of label l_i in the discrete label distribution corresponding to the block m of the kth training volume. Similarly, we can obtain the relation matrix $\Phi_y^{mn}(i, j)$, $\Phi_z^{mn}(i, j)$ for the y and z direction, respectively.

With the MRF model, the probability of a spatially or temporally adjacent label pair can be defined as follows:

$$p(< l_i, l_j >; \phi_\alpha^{mn}) = \frac{1}{Z(\phi_\alpha^{mn})} \exp(\phi_\alpha^{mn}(i,j)) \tag{12}$$

where $Z(.)$ is the partition function, $\phi_\alpha^{mn}(i, j)$ is the interaction potentials which are learned from the training data, α stands for the category of relation that is one of x, y, z or t, m and n denote the spatial positions of two motion patterns where l_i and l_j are extracted. Using the joint probability as the weight, we compute the consistent probability of the two adjacent motion patterns in a weighted sum form as follows:

$$p(< M_m, M_n >; \phi_\alpha^{mn}) = \sum_{i=1}^{|\mathcal{L}_{M_m}|} \sum_{j=1}^{|\mathcal{L}_{M_m}|} p^{M_m}(l_i) p^{M_n}(l_j) p(< l_i, l_j >; \phi_\alpha^{mn}) \tag{13}$$

where M_m and M_n are two motion patterns located at spatial position m and n, respectively. $p^{M_m}(.)$ and $p^{M_n}(.)$ are the discrete distributions corresponding to the motion pattern pair, ϕ_α^{mn} is the parameter matrix of the interaction potential learned from the training data.

Thus, given an input labeled dataset $\mathcal{D}$, the likelihood function for relation R^{mn} is calculated by

$$p(D; \phi_\alpha^{mn}) = \frac{1}{Z(\phi_\alpha^{mn})^{|\mathcal{D}|}} \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \Phi_\alpha^{mn}(i,j) \phi_\alpha^{mn}(i,j) \tag{14}$$

where $\Phi_\alpha^{mn}(m, n)$ is the entry (i, j) of the frequency matrix for spatial relation R^{mn} learned from the training data set $\mathcal{D}$, which counts the virtual times that the label pair (l_i, l_j) appears in a training lattices at the spatial position m and n, respectively. $|\mathcal{D}|$ is the total number of the volumes in the training set $\mathcal{D}$.

With the spatio-temporal context modeling that captures the co-occurrence relations among adjacent motion patterns, we define a measurement that indicates the consistency between a specific motion pattern and its spatial and temporal neighbors:

$$C_{context}(M_m) = \min_{\beta \in neighbors} p(< M_m, M_\beta >) \tag{15}$$

where β is the adjacent spatio-temporal position, which has one of six relations relative to the current position m - spatial up, spatial down, spatial left, spatial right, spatial before, spatial after, temporal before or temporal after.

V. ANOMALY DETECTION

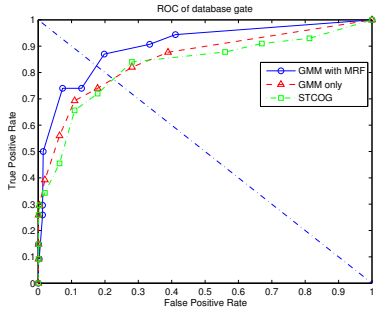
With the given discussions, we can now integrate the static and the contextual measurements for detecting anomaly motions in a crowded scene. Using the spatio-temporal context learned from training data, we evaluate the confidence measure for each usual motion pattern, simultaneously taking into account the confidence C_{basic} that indicates the consistency with the particular prototype, and the confidence $C_{context}$ that denotes the consistency of the spatio-temporal relations with the neighbors in the scene. Specially, we use a weighted combination of the two factors as follows:

$$C_{total} = \beta C_{basic} + (1 - \beta) C_{context} \tag{16}$$

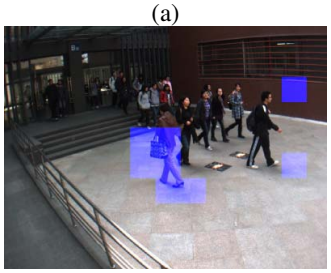
In this way, we finally use the summarized confidence C_{total} to detect anomaly events from an input image sequences of a crowded scene. When the total confidence is lower than a specific threshold, we consider the motion as an anomaly event.

TABLE III. PERFORMANCE COMPARISONS OF THE PROPOSED METHOD BY BOTH GMM AND MRF, GMM ONLY, AND THE STCOG APPROACH [16], IN WHICH TP AND FP DENOTE THE TRUE POSITIVE AND THE FALSE POSITIVE, RESPECTIVELY).

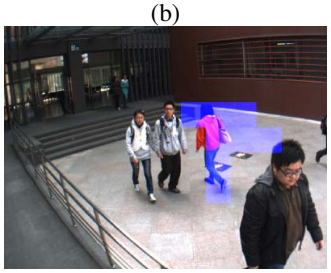
Crowded Scene Categories	GMM with MRF		GMM only		STCOG [16]	
	True Pos	False Pos	True Pos	False Pos	True Pos	False Pos
Outdoor *Building Gate*	0.81%	0.19%	0.78%	0.22%	0.78%	0.22%
Indoor *Classroom*	0.80%	0.19%	0.72%	0.27%	0.72%	0.27%
Outdoor *Street*	0.68%	0.31%	0.59%	0.42%	0.58%	0.43%



(a)

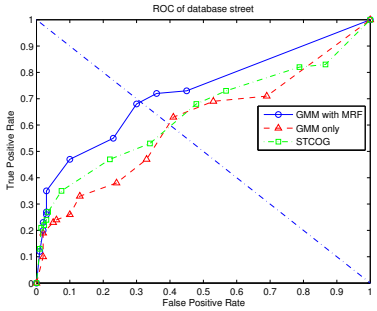


(b)

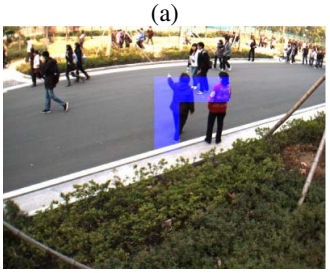


(c)

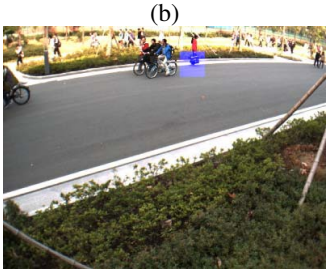
Fig. 2. Outdoor abnormal detection (*Building Gate*): (a) the ROC curves of the proposed method by both GMM and MRF, GMM only, and the STCOG approach, (b) and (c) the detected abnormal events of *walking along a different direction*.



(a)



(b)



(c)

Fig. 3. Outdoor abnormal detection (*Street*): (a) the ROC curves of the proposed method by both GMM and MRF, GMM only, and STCOG, (b) the detected abnormal event of *suddenly stopping*, and (c) the detected abnormal event of *changing direction*.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

We evaluate the performance of our anomaly detection on the dataset [15], which consists of over 25000 images with the resolution of 1024×768 pixels. The dataset contains samples of three crowded environments: two outdoor scenes, namely, a (i) *Building Gate* scene, a (ii) *Street* scene, and an indoor *Classroom* scene. Each scene category contains several abnormal cases such as *stopping*, *turning around*, *suddenly changing speed* and *going backwards*. For each abnormal event, we averagely collect 250 images for testing, while leaving about 1500 images for training. To evaluate the effectiveness of our anomaly detection method, all the query sequences are hand-labeled and used as the ground truth. Note that [15] presented a model for abnormal detection from depth videos, in which context modeling is not considered.

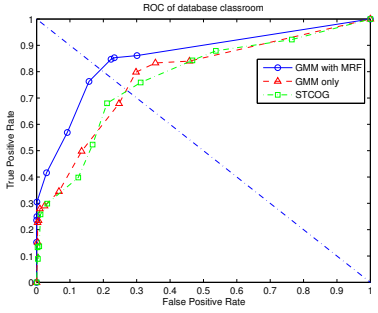
Table III shows the comparison results of three different approaches on the same dataset: the proposed algorithm with both GMM and MRF modeling, the proposed algorithm with GMM only, and another approach STCOG [16]. STCOG is also a block-level algorithm for detecting abnormal events, in which GMM is built on every two neighboring blocks. It can be found that the performance of the proposed algorithm with GMM only is very similar to those of STCOG. However, after context modeling using MRF, the average true positive rate of our method is increased while the average false positive rate is decreased simultaneously.

Fig. 2 shows the ROC curves of the three methods and several detected abnormal event examples from an outdoor scene of *Building Gate*. The ROC curves in Fig. 2(a) illustrates the effectiveness of anomaly detecting by combining GMM

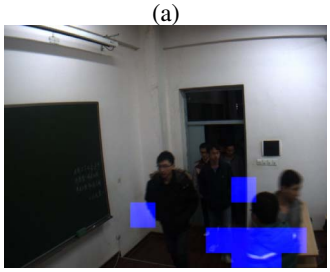
and MRF modeling in such a crowded scene. Fig. 2(b) and Fig. 2(c) show two detected cases of the abnormal event *walking along a different direction*, respectively.

Fig. 3(a) further shows the ROC curves of the three methods in another outdoor *Street* scene. Similarly, Fig. 3(b) and Fig. 3(b) give two examples of the detected anomaly motions. The rest abnormal events, such as *unexpected turning back*, *sudden speed changes*, and *walking along wrong directions* can all be successfully detected. Note that the ROC curves of the *Street* scene in Fig. 3(a) are lower than those in Fig. 2(a). It is due to the fact that the motions in a *Street* scene is in general much more complex than the latter. For instance, pedestrians and vehicles such as cars and bicycles may simultaneously exist in the same scene, making it difficult to clearly define the steady motion patterns. Moreover, pedestrians become smaller in such a street scene, which potentially makes the detection of their anomaly motions easily being disturbed.

Fig. 4 gives the comparisons using the image sequences from an indoor *Classroom* scene. For such an indoor scene, we define the motions of entering the classroom as usual events. Generally, the accurate calculation of optical flows plays an important role in our GMM and MRF integrated anomaly detection. We find there may exist errors during calculating optical flows in an indoor scene, potentially brought by the low-quality of illumination or the easy confusion of scene objects and the scene background. Fig. 4(b) shows the correctly detected anomaly motions together with a false positive which is caused by inaccurate optical flow estimation.



(a)



(b)

Fig. 4. Indoor scene detection: (a) the ROC curves of the Classroom scene the proposed method by both GMM and MRF, GMM only, and STCOG, and (b) the detected abnormal motion in the indoor *Classroom* scene.

VII. Conclusion

A novel statistical framework for modeling the intrinsic structure of crowded scenes and detecting abnormal activities is presented in this paper. We propose an atomic motion pattern representation using the Gaussian Mixture Model to distinguish the motions inside each block in a refined way. Usual motion patterns can thus be defined as a certain type of steady motion activities appearing at specific scene positions. We further use the Markov Random Field model to characterize the joint label distributions over all the adjacent local motion patterns inside the same crowded scene, aiming at modeling the severely occluded situations in crowded scene accurately. By combining the determinations from the two stages, we propose a weighted scheme to automatically detect anomaly events from crowded scenes. The experimental results on several different outdoor and indoor crowded scenes illustrate the effectiveness of the proposed algorithm. In the future, we will focus on incorporating stronger scene context information to further improve the performance, and improve the robust calculation of optical flows.

Acknowledgment

The work described in this paper was supported by the Natural Science Foundation of China under Grant No. 61272218 and No. 61321491, the 973 Program of China under Grant No. 2010CB327903, and the Program for New Century Excellent Talents under NCET-11-0232.

References

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

[2] D. Kong, D. Gray and H. Tao, *Counting pedestrians in crowds using viewpoint invariant training*, BMVC, pp. 1-6, 2005.

[3] S. Ali, and M. Shah, *Floor fields for tracking in high dendity crowd scenes*, ECCV, pp. 1-14, 2008.

[4] B.T. Morris, and M.M. Trivedi, *A survey of vision-based trajectory learning and analysis for surveillance*, IEEE Transactions on Circuits and Systems for Video Tehcnology, vol. 18, no. 8, pp. 1114-1127, 2008.

[5] W. Hu, D. Xie and T. Tan, *A system for learning statistical motion patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1450-1464, 2004.

[6] T. Zhao, and R. Nevatia, *Tracking multiple humans in crowded environment*, CVPR, pp. 406-413, 2004.

[7] W.M. Hu, X.J. Xiao, Z.Y. Fu, D. Xie, T.N. Tie, and S. Maybank, *A system for learning statistical motion patterns*, IEEE Trans.on Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1450-1464, 2006.

[8] E.L. Andrade, S. Blunsden and R.B. Fisher, *Modelling crowd scenes for event detection*, ICPR, pp. 175-178, 2006.

[9] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, *Robust real-time unusual event detection using multiple fixed-location monitors*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, no. 3, pp. 555-560, 2008.

[10] V. Mahadevan, W.X. Li, V. Bhalodia, and N. Vasconcelos, *Anomaly detection in crowded scenes*, CVPR, pp. 1975-1981, 2010.

[11] T. Xiang, and S. Gong, *Video behaviour profiling and abnormality detection without manual labelling*, ICCV, pp. 1238-1245, 2005.

[12] B. Zhao, F.F. Li, and E.P. Xing, *Online detection of unusual events in videos via dynamic sparse coding*, CVPR, pp. 3313-3320, 2011.

[13] L. Kratz, and K. Nishino, *Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models*, CVPR, pp. 1446-1453, 2009.

[14] C. Galleguillos, A. Rabinovich, and S. Belongie, *Object categorization using co-occurrence, location and appearance*, CVPR, pp. 1-8, 2008.

[15] X.L. Ma, T. Lu, F.M. Xu, and F. Su, *Anomaly Detection with Spatio-Temporal Context Using Depth Images*, ICPR, pp. 2590-2593, 2012.

[16] Y. Shi, Y. Gao, and R. Wang, *Real-time abnormal event detection in complicated scenes*, ICPR, pp. 3653-3656, 2010.