Abnormal Event Detection Based on Multi-scale Markov Random Field

Lei Qin[1](✉), Yituo Ye[2], Li Su[2], and Qingming Huang[2]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
qinlei@ict.ac.cn
[2] School of Computer and Control Engineering,
University of Chinese Academy of Sciences, Beijing 100190, China

Abstract. In this paper, we present a novel unsupervised method for abnormal behavior detection, which considers both local and global contextual information. For the local contextual representation, we firstly divide video frames into local regions, then extract low-level feature such as histogram of orientated optical flow (HOF) and sequential feature which is composed of K temporal adjacent frames for each region. The global contextual feature encodes the statistical characteristics of those local features like orientation entropy and magnitude variance. An online clustering algorithm is introduced to generate dictionaries for the local and global features respectively. Then, for any new incoming feature, a maximum posterior estimation of the degree of normality is computed by multi-scale Markov Random Field (mMRF) based on the learned model. The proposed method is evaluated on hours of real world surveillance videos. Experimental results validate the effectiveness of the method, and the detection performance is promising.

Keywords: Computer vision · Anomaly detection · Multi-scale markov random field

1 Introduction

Detecting abnormal behaviors in videos is one of the most promising fields in computer vision. It is receiving increasing attention due to its wide range of practical applications such as smart surveillance, suggesting frames of interest that should be analyzed by an expert, and summarizing the interesting content.

However, there are still several problems in anomaly detection especially for the scene consisting of complex correlated activities performed by multiple people. Firstly, unusual activities seldom occur and the large intraclass diversity of unusual and usual activities makes them even harder to be predefined. The main paradigm for abnormality detection in videos recently is to extract features and to learn a model on normal samples from the video. So that anomaly is detected as the one fitting the model badly. Various methods may differ in the feature they used and the model they built. Secondly, the visual context for scene

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 376–386, 2015.
DOI: 10.1007/978-3-662-48558-3_38

tends to change over time, which makes the incrementally updated process even more necessary.

Based on these problems, several methods have been proposed. Specifically,[2], [4], [11] determine abnormality based on the trajectory for each object. However, trajectory is too dependent on the tracking algorithm and may be unreliable in crowd scenes. [1] proposes a simple approach that measures typical optical flow speed and direction for each local grid to determine anomaly. Yet this method discards the relationship among local regions which may contain the contextual information. Approaches using Bayesian topic model [7], [13] evaluate the abnormality based on the interaction of local activities, but they only run in batch mode. Mehran et al. [9] present a new way to formulate the abnormal crowd behavior by adopting the social force model, and then use Latent Dirichlet Allocation (LDA) to detect abnormality. In [12], they define a chaotic invariant to describe the event. [6] utilizes a space-time Markov random field model for abnormality detection and the events that could not be described by the model is regarded as anomaly. [14] provides a framework using sparse coding method which builds a dictionary for spatial-temporal cuboid dynamically and anomaly is detected as the one with a large value of the proposed objective function.

Methods[9] can be viewed as “global”, for they attempt to find the global abnormal event (GAE) in a video clip. Methods [1], [14], [6], in contrast, focus on the local abnormal event (LAE). However, few methods can be applied both in global and local scale as the situations are often different. Additionally, most methods directly utilize the low level feature such as optical flow, but these features may be not stable enough and may discard some useful contextual information. So how to take contextual information into consideration is important for the abnormal event detection.

In this work, we introduce a novel unsupervised method based on contextual information and mMRF. In the feature part, both local and global features are utilized, which are corresponding to the different scales of the mMRF. For the local scale, we divide each frame of the video into a grid of local regions. HOF is used to encode the low level information, and sequence composed of K adjacent frames is also extracted for the local feature representation. Sequence characters the temporal relationship of the low level features for each local region and it may bring more contextual information. The global feature corresponds to the global scale in mMRF, and it is used to describe the situation for the whole frame with utilizing features such as orientation entropy,magnitude variance. In the model part, mMRF is employed. It can describe the spatial relationship between the local features. Different from the space-time markov random field model proposed by [6], structure of mMRF is hierarchical and it combined different scale of features. It can deal with both GAE and LAE.

The contributions of our methods are mainly two folds: first, contextual information is introduced to describe the action more precisely, take loitering activity in the subway station dataset [1] as an example, a person would stay a few frames in the video and this may be well described by the sequential information.

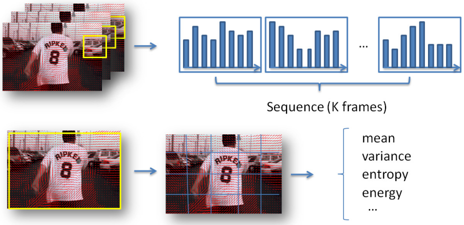


Fig. 1. Both local and global feature, the upper row show the sequential information used in this paper, and the bottom one show the global feature which is the statistical information of all the local feature.

Second, mMRF extends MRF to multiple scale, it combines the features in different scales so that mMRF can cope with both LAE and GAE.

2 Activity Representation

In this section, local feature, global feature and the corresponding similarity function are presented in section 2.1. Then model acquisition and maintenance using online clustering algorithm is illustrated in section 2.2.

2.1 Features and the Corresponding Similarity

Local Feature. First, each frame is divided into M by N local regions. The number of local regions depends on how finely we want to capture the motion details. For every local region, two kinds of information are utilized, the low level information (HOF) and the sequential information.

As Fig.1 shows, a sequence is defined as K temporal adjacent frames for every local region, which can be represented by K histograms. Sequence takes the temporal relationship between the HOF features into consideration and it may bring more contextual information. Over all, we combine HOF and sequential information to describe the local activity in this paper.

Similarity of the Local Feature. For the low level information such as HOF, we directly use the common chi-square distance to measure their similarity, and we denote $sim_f(f_1, f_2)$ for the similarity of these features, where f_1 and f_2 denote the low level feature. As for the sequential information proposed above, the similarity of sequences should obey several requirements. 1) The similarity should take alignment into consideration. For sequences of the same action may be segmented in different ways but their similarity should be high. 2) The similarity should be able to measure sequences with different length.

We utilize the edit distance [10] mainly used in natural language processing to measure the similarity between sequences. On the whole, edit distance measures the number of operations required to transform a string into another. The basic operations include replacement, delete and insert. For example, edit distance of string ‘1234’ and ‘123’ is 1 with a delete or insert operation, and distance of string '1234' and '1235' is also 1 with a replace operation. Meanwhile, edit distance of two strings sharing a similar structure would be small. Edit distance of string '1234' with '4123' is 2, with a delete and insert operation, however the traditional distance of them is 4, for every two elements at the same position are different. So it can contribute to the alignment. As described above, requirement 1) and 2) would be satisfied by the edit distance. We denote $sim_s(s_1, s_2)$ as the similarity of sequences. s_1 and s_2 just represent two sequences.

It should be noted that elements of the sequence are histograms. Chi-square distance and a threshold θ is used to determine whether they are equal in this paper, given by

$$equal(H_1, H_2) = \begin{cases} 1, \chi^2(H_1, H_2) < \theta \\ 0, otherwise \end{cases} \tag{1}$$

Based on this function, edit distance [10] can be introduced to measure the similarity between the two sequences and it is computed using a dynamic programming algorithm.

Global Feature. Global feature should represent the condition of the whole frame, and we use statistical value of the local regions to characterize it, for example, the mean moving orientation, the disorder of the orientation and so on. In this paper, orientation entropy, magnitude entropy, orientation variance, magnitude variance [5] and Kinetic Energy [15] are utilized. By using these global features, we can have a general idea about what happened in each frame, and this may be hard for local feature alone. For the distance for global feature, we just use the Euclidean distance to measure their similarity and $sim_g(g_1, g_2)$ is used to denote the similarity, where g_1 and g_2 denote the global feature.

2.2 Model Acquisition and Maintenance

For the first M frames, with the local feature as sequence and the global feature, an online clustering algorithm is introduced to construct the model. In this paper, basic leader-follower clustering algorithm [3] is utilized. The main procedure is as follows: given a new sample x, find its nearest clustering center w_j and the corresponding distance d. If d is smaller than a threshold σ , the clustering center should be modulated, otherwise a new clustering center should be added. Dictionaries of the local feature and global feature are stored in the model, and they would be updated by the new coming features.

3 Abnormality Detection Based on Multi-scale MRF

3.1 Structure of Multi-scale MRF

In this paper, we just use two scales for the mMRF, global and local scale. As the Fig. 2 shows, the global scale is for the full frame, and we divide frames into a grid of small regions (M by N) for local scale. Each local region denotes a local node and the whole frame represents the global node. The blue nodes mean the

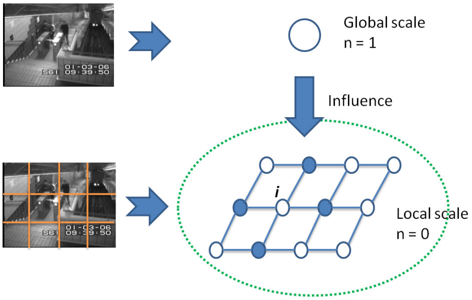


Fig. 2. Structure of the mMRF.

neighbors of node i in the local scale. The global node may affect all the local nodes, which may act as some kind of prior knowledge. Then for node i, its state may be determined by its similarity with the model, its neighbors in local scale and the global node in the global scale. We combine the different scale by the energy function. Based on the centers and the corresponding frequencies for the features, we can get the node evidence and the pairwise evidence. Ultimately, inference on the graph will yield the maximum a posteriori (MAP) labeling that specifies which nodes are normal or abnormal.

3.2 Energy Function of the Multi-scale Markov Random Field

The energy function in the mMRF model is following:

$$E(X) = \sum_i \gamma_i E_i(X) \tag{2}$$

where i denotes different scales and γ_i is the weight for each scale. As for this paper, $E(X) = \gamma E_{local}(X) + (1 - \gamma) E_{global}(X)$, where $E_{local}(X)$ denotes the local scale energy function and E_{global} denotes the global scale energy function. γ is used to weight the two scales.

For the local scale, the energy consists of two parts: node evidence and pairwise evidence and it can be represented as:

$$E_{local}(X) = \sum_i h(x_i) + \alpha \sum_{i,j \in neighbour} s(x_i, x_j) \tag{3}$$

where $h(x_i)$ is the node evidence and $s(x_i, x_j)$ is the pair-wise evidence. The value α is a constant to weight the pair-wise evidence, and x_i denotes the label telling whether node i is normal or abnormal. ($x_i = 0$ signifies node i is normal and $x_i = 1$ signifies node i is abnormal).

The node evidence function measures the similarity between the event and the model. It can be divided into two terms: the similarity for the HOF feature $h_f(x_i)$ and the similarity for the sequential feature $h_s(x_i)$. Simply speaking, for node i, if the HOF feature and sequential feature are very similar with the clustering center always occurring before, $h_f(x_i = 0)$ and $h_s(x_i = 0)$ will become higher.

Complementarily, $h_f(x_i = 1) = 1 - h_f(x_i = 0)$, $h_s(x_i = 1) = 1 - h_s(x_i = 0)$. We compute both $h_f(x_i = 0)$ and $h_s(x_i = 0)$ based on the model:

$$\begin{cases} h_f(x_i = 0) = \sum_j \sum_k fref_k \times sim_f(f_{i,j}, mf_k) \\ h_s(x_i = 0) = \sum_j fref_j \times sim_s(s_i, ms_j) \end{cases} \tag{4}$$

where $fref_k$ and $fref_j$ denote the frequency of HOF clustering center mf_k and sequence feature clustering center mf_j, it is defined as the possibility of the occurrence of the feature cluster, which is computed as the number of the samples of each clustering center divided by the total samples. $f_{i,j}$ denotes the jth HOF for observation, s_i denotes the sequential information, $sim_f(\cdot)$ is the similarity for HOF and $sim_s(\cdot)$ denotes the sequence similarity proposed in section 2. Both $h_f(x_i = 0)$ and $h_s(x_i = 0)$ are computed as the sum of the product of the clusters frequency and the similarity of the feature and the clustering center. Abnormal events seldom happen, and their similarities with most clustering centers are small. When abnormal event happened, $h_f(x_i = 0)$ and $h_s(x_i = 0)$ may have a small value then $h_f(x_i = 1)$ and $h_s(x_i = 1)$ will be high. We combine $h_f(x_i)$ and $h_s(x_i)$ for the node evidence, and $h(x_i = 0) = (1 - \tau)h_f(x_i = 0) + \tau h_s(x_i = 0)$, where τ is a weighting constant set with $0 < \tau < 1$.

The pair-wise evidence function measures the similarity between the neighboring nodes. When $x_i = 0$, $x_j = 0$, $s(x_i = 0, x_j = 0) = sim_s(s_i, s_j)$, and $s(x_i = 1, x_j = 1)$ is also defined as $sim_s(s_i, s_j)$. Otherwise, $ss(x_i, x_j) = 1 - s(x_i = 0, x_j = 0)$. For the global scale, the energy is set as $E_{global}(X) = \sum_j \sum_k freg_k \times sim_g(g_j, mg_k)$, where g denotes the coming global feature, $freg_k$ denotes the frequency of global feature clustering center mg_k. The same as the local energy function, when the event is abnormal, $E_{global}(X = 1)$ should have a large value and $E_{global}(X = 0)$ would be high in normal condition.

Given the parameters for every node and link of the mMRF, we carry out MAP inference to maximize the energy function. Loopy belief propagation with max-sum message passing is used, which provides the MAP labeling whether each node is normal or not.

4 Experimental Results

In this section, we show the empirical performance of the proposed abnormal event detection algorithm on several published datasets. Section 4.1 introduces global abnormal event detection based on the UMN dataset[1]. Experiments on local abnormal event detection are introduced in section 4.2, which is based on the subway station dataset provided by Adam et al.[1]. The cross validation strategy is used to select parameters.

[1] Unusual crowd activity dataset of the University of Minnesota. (http://mha.cs.umn.edu/movies/crowdactivityall.avi).

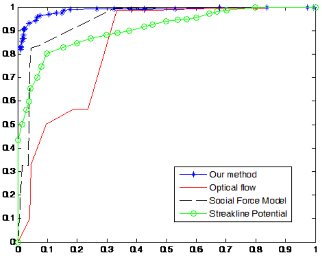


Fig. 3. The ROCs for frame level GAE detection in the UMN dataset.

Table 1. Comparison of the accuracy by our approach and other methods.

Method	AUC
Social Force [9]	0.96
Optical flow	0.84
Streakline Potential [8]	0.90
Ours(mMRF)	0.973
Ours (HOF alone)	0.624

4.1 Global Abnormal Event Detection

Datasets. We use the UMN dataset to verify the effectiveness of our method on the GAE. The UMN dataset consists of 11 clips of the crowded escape events in 3 different scenes including both indoor and outdoor scenes.

Experimental Results. For every clip, we use the first 400 frames for training and rest for testing (We use the first 250 frames for training in clip 3, for abnormal has already happened when it comes to the 400th frame). The local region size is 16×16 for each one, and we just use HOF to represent the local feature. Because the abnormal in UMN is GAE, the parameter γ should be set low to add weight of the global feature. In this experiment, γ is set to be 0.2, and the other parameter are set as $\alpha = 0.6$, $\tau = 0.5$. We use different threshold θ for the online clustering algorithm to get the ROC curve as showed in Fig.3, and value in the curve is set to be the mean of all the 11 clips. We compare the results using mMRF which combined all the features and the method only using local feature alone.

From the ROC curve shown in Fig.3 and Table 1, we can see that the method using mMRF perform better, the mean AUC of methods using mMRF is 0.973 and it outperforms 0.624 using local feature. The reason mainly due to the contribution of the global feature. People are just wandering in normal condition. When abnormal happened, they escaping all round or just in one direction. Only from a fixed local region, it may be hard to determine whether abnormal happens, but combined with the global feature such as kinetic energy, the detection

Table 2. Comparison of the accuracy using HOF information alone and other sequence similarity. Numbers in the first row denotes count for each abnormal activity in the ground truth.

	LT	NP	WD	II	Misc	Total	False alarm
Ground truth	14	13	26	4	9	66	-
Ours(HOF)	8	7	24	2	2	43	23
Ours(HOF + chi-square distance)	13	8	23	4	8	56	15
Ours(HOF + edit distance)	14	8	24	4	8	58	5
Jaechual Kim[6]	13	8	24	4	8	57	6
Bin Zhao[14]	14	9	25	4	8	60	5

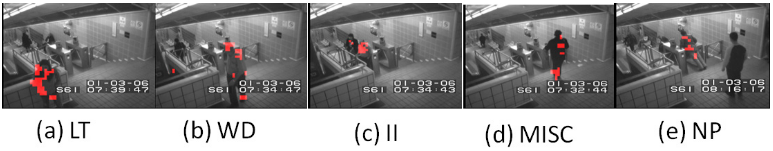


Fig. 4. Examples of the detected unusual event in the subway entrance surveillance video by our algorithm. LT: loitering; WD: wrong direction; NP: no payment; MISC: misc; II: irregular interactions between people.

may be much easier. Moreover, we also provide the quantitative comparisons to the state-of-the-art methods, the AUC of the method using mMRF range from 0.951 to 0.985, and it is comparable with the method in [9] for 0.96 and the method [8] for 0.90.

4.2 Subway Station Dataset

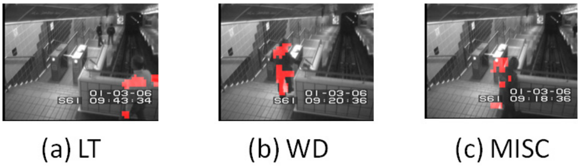
The dataset used for LAE are two video sequences taken from a fixed surveillance camera at a subway station, one monitoring the exit gate and the other monitoring the entrance gate. In both cases, there are one to ten people appearing in the scene at the same time. The frame size is 512×384, and the length of videos are 96 and 43 minutes correspondingly.

For both videos, we divide every frame into 64×48 local regions and extracts HOF and sequential feature from each region. Sequence length in this paper is set as 10. Global feature described in section 2 is also used. On the whole, two kinds of experiments are carried out. First, we verify the effectiveness of the sequential information we proposed. As a comparison, anomaly detection which utilizes the HOF alone is also conducted. Then we do experiments to verify the good performance of our sequence similarity. Sequence similarity in this paper is based on the edit distance, and chi-square distance is used for comparison.

The Entrance Gate. For the entrance gate, we use the first 12 minutes for training and rest for testing. The local region size is 8×8 for each one, and the length of the sequence is set to be 10. The parameter γ was set 0.8 so that the weight of the local feature is high. And we set $\alpha = 0.6$, $\tau = 0.5$, $\theta = 1.4$ correspondingly. Because of the stationary camera, we conducted background

Table 3. Comparison of the accuracy using HOF information alone and other sequence similarity. Numbers in the first row denotes count for each abnormal activity in the ground truth

	LT	WD	Misc	Total	False alarm
Ground truth	3	9	7	19	-
Ours(HOF)	2	9	2	13	14
Ours(HOF + chi-square distance)	3	8	7	19	9
Ours(HOF + edit distance)	3	9	7	19	3
Jaechual Kim[6]	3	9	7	19	3
Bin Zhao[14]	3	9	7	19	2



(a) LT (b) WD (c) MISC

Fig. 5. Examples of the detected unusual event in the subway entrance surveillance video. LT: loitering; WD: wrong direction; MISC: misc; MISS: miss; FA: false alarm.

subtraction first and extracted features from the foreground. The results of the experiment are as Table 2. Fig.4(a) - (e) show the examples for abnormal activities LT, WD, II, Misc and NP correspondingly.

The second row and the fourth row of Table 2 show the comparison of the anomaly detection with and without using the sequential information. Both of them provide similar results in the abnormalities such as "wrong direction". For the feature HOF is quite useful in describing the motion direction information. However, the method which does not utilize the sequence performs poorly in the abnormalities such as "Misc" which is often caused when a person abruptly stops walking or runs fast. The reason may be that HOF alone discards the temporal information between the low level features. Besides, false alarm rate for this method is high because optical flow information is not stable and is sensitive to the optical flow parameters and illumination changes. The third row and the fourth row compare with the results using chi-square distance and edit distance. As analyzed before, chi-square distance does not take alignment into consideration and it may bring a relatively high false alarm rate. We also compare our results with the method used in [6] and [14] in the fifth and the sixth row, and we can see that the results are comparable. It should be noted that, the method in [6] and [14] is specially designed for the local abnormal event detection and it may be not suitable for the global abnormal event detection.

The Exit Gate. For the exit gate, we use the first 8 minutes for training and rest for testing. The other parameters are set the same as the entrance gate. The results can be seen in Table 3. Fig.5(a) - (c) are detected by our methods, which corresponds to the LT, WD and Misc.

Same as the entrance gate, the second row and the fourth row of Table 3 show the comparison of the anomaly detection with and without using the sequential information. And the third row and the fourth row compare with the results using chi-square distance and edit distance. The results are consistent with the Table 2 and these results may verify the effectiveness of our methods.

5 Conclusion

In this paper, we propose a novel unsupervised framework based on contextual information and multi-scale markov random field for abnormal behavior detection. Both local and global features are utilized, and each corresponds to different scales of the multi-scale markov random field. With combing these features in mMRF, both GAE and LAE can be detected, and the experimental results verify the effectiveness of the proposed method.

Acknowledgments. This work was supported in part by National Basic Research Program of China(973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61332016, 61133003, 61390510, 61472389.

References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 555–560 (2008)
2. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification and scene analysis (1995)
4. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A system for learning statistical motion patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9), 1450–1464 (2006)
5. Ihaddadene, N., Djeraba, C.: Real-time crowd motion analysis. In: 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
6. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2928. IEEE (2009)
7. Li, J., Gong, S., Xiang, T.: Global behaviour inference using probabilistic latent semantic analysis. In: The British Machine Vision Conference (2008)
8. Mehran, R., Moore, B.E., Shah, M.: A streakline representation of flow in crowded scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 439–452. Springer, Heidelberg (2010)
9. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 935–942. IEEE (2009)
10. Navarro, G.: A guided tour to approximate string matching. ACM computing surveys (CSUR) **33**(1), 31–88 (2001)

11. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8), 747–757 (2000)
12. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2054–2060. IEEE (2010)
13. Xiang, T., Gong, S.: Incremental and adaptive abnormal behaviour detection. Computer Vision and Image Understanding **111**(1), 59–73 (2008)
14. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3313–3320. IEEE (2011)
15. Zhong, Z., Ye, W., Wang, S., Yang, M., Xu, Y.: Crowd energy and feature analysis. In: IEEE International Conference on Integration Technology, pp. 144–150. IEEE (2007)