International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.4 (2014), pp.115-126
http://dx.doi.org/10.14257/ijsip.2014.7.4.11

Abnormal Event Detection in Nature Settings

Yang Liu[1], Yibo Li[2] and Xiaofei Ji[2]

[1]College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]College of Automation, Shenyang Aerospace University, Shenyang, China
Yang97_net@163.com

Abstract

Abnormal event detection in nature settings is an active issue in computer vision domain. A novel unsupervised method is proposed to detect abnormal events by combining dynamic texture and sparse coding. In this method, dynamic texture is used as descriptors in a spatio-temporal manner to describe spatio-temporal volumes of events in videos. Sparse coding is utilized for reconstructing the testing data to measure its normalness. Experiments are conducted on the well known UCSD dataset and UMN dataset to demonstrate the efficiency of the proposed method. The results show that the proposed method outperforms the current state-of-the-art methods.

Keywords: *computer vision, abnormal event detection, sparse coding, dynamic texture*

1. Introduction

During the last few years, video surveillance is one of the most important applications of computer vision. Of the many possible tasks, detecting abnormal events from video sequence is of considerable practical importance. The developments have been achieved and many methods have been implemented to carry out automatic abnormal event detection. A structured and comprehensive overview of the research can be viewed in [1]. Anomaly detection techniques can operate in three protocols, *i.e.* supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection. Techniques developed in a supervised model [2-4] assume a training dataset is available, in which the data are labeled for both normal cases and anomalies cases [1]. The main advantage of the supervised method clearly lies in the fact that the normal and abnormal models are known precisely. For example, Nasution and Emmanuel [4] define all the normal as standing, sitting, bending/squatting, and abnormal events as side lying and lying before detection. The disadvantage lies in the fact that it is impossible to obtain all the normal and abnormal models before detection. Semi-Supervised anomaly detection assumes that the training data is labeled only for normal cases. Since techniques that operate in a semi-supervised mode do not require labels for the anomaly class, their applications are more widely than supervised techniques. The major limit is that it is difficult to obtain a training data set which covers every possible normal behavior that can occur in the data. Unsupervised detection [5-7] does not require training data, and it is the most widely used in nature settings. Without prior training data, a database of normal events is build up in a completely unsupervised manner to detect deviation [5]. The techniques in this category make the implicit assumption that normal instances are far more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from high false alarm rate.

ISSN: 2005-4254 IJSIP

Inspired by these works, a novel method of combining dynamic texture (DT) and sparse coding is proposed for detecting abnormal events in videos. The proposed method is completely unsupervised, making no prior assumptions of what abnormal events may look like. The unsupervised manner is based on the assumption that abnormal events occur rarely, normal events occur often and are obtained easily in the initial time.

Abnormal event detection consists of two important aspects: video representation and abnormal event detection. Currently, the most popular proposal video representation is the spatio-temporal interest point [8]. First, the Gaussian filter is used in the space plane and the Gabor filter is used in the temporal plane to obtain interest point cuboids. Then, the cuboids are described by descriptors, such as flattening all the pixels, histogram of optical flow or gradients. This type of computation is expensive because the number of cuboids detected is in the thousands. To address this problem, a novel method is developed to describe cuboids to detect abnormal event in videos. The proposed method uses dynamic texture descriptor, Local Binary Patterns by Three Orthogonal Planes (LBP-TOP), to describe cuboids in a spatio-temporal way. The LBP-TOP feature has been successfully used for facial expression recognition [9]. Vili Kellokumpu [10] use xt and yt slices as features for human action recognition, by which the conclusion can be drawn that the LBP-TOP has discriminative ability to describe spatio-temporal interest points. In the proposed method, the LBP-TOP is used as descriptor of cuboids. For the other aspect of abnormal event detection, sparse coding is utilized. Sparse coding is a popular classification method in recent years. An over-complete dictionary is defined or trained by user. Then a coefficient vector is obtained by using iterative algorithm to realize mapping from high-dimensional data space to a low-dimensional space. The coefficient is sparse and can characterize the main features of the data. Therefore, the test data can be reconstructed by using a little number of bases in the dictionary. In 2009, Wright and Yang, *et al.*, [11] verified that the sparse coding demonstrated good performance in face recognition. The method can guarantee good recognition rate, though a lot of noise and occlusion exist. Then, the sparse coding was introduced into the field of pattern recognition and semantic understanding, and many scholars have conducted in-depth research and development. It has been successfully applied in various research areas, such as face recognition [12, 13], image classification [14, 15], action recognition [16-18], detection of abnormal events [19, 20], background subtraction [21] and tracking [22]. The main trait of sparse coding is that the observed data can be constructed by linear combination of sparse number of basis of the over-complete dictionary. As we know it is often true that only a small portion of video contains important information, so sparse coding can be used to address the problem of the huge amount of data in abnormal event detection.

In summary, the LBP-TOP descriptor combined with the sparse coding is proposed as a novel method for abnormal event detection. This approach is developed specifically for the unsupervised anomaly detection protocol, and assumes that the normal events can be obtained in the initial time of the detection procedure. To the best of our knowledge, the approach of the combination of the dynamic texture and sparse coding to realize abnormal event detection has not been presented in any literature. Experiments show that the proposed method outperforms the current state-of-the-art methods.

The remainder of this manuscript is organized as follows. Section 2 provides detailed explanation of the dynamic texture, followed by a brief review of previous works on sparse coding and the abnormal event detection method in Section 3. Section 4 demonstrates the effectiveness of the proposed method using real world surveillance videos, followed by conclusions in Section 5.

2. Video Representation

The task of video representation includes two steps: interest points detection and the descriptor. The proposed abnormal event detection algorithm adopts a representation based on spatio-temporal cuboids. The method [8] is used to detect salient points within the video and LBP-TOP descriptor is used to describe the local spatio-temporal cuboids around the detected interest points.

2.1. Spatio-Temporal Cuboids

The interest points proposed [8] is widely used in video analysis. At each interest point, a threshold is set for the local maximum of the response function. The points, where the response values of filters are bigger than the threshold, are extracted. Figure 1 shows spatio-temporal interest points detected by using the method in [8].

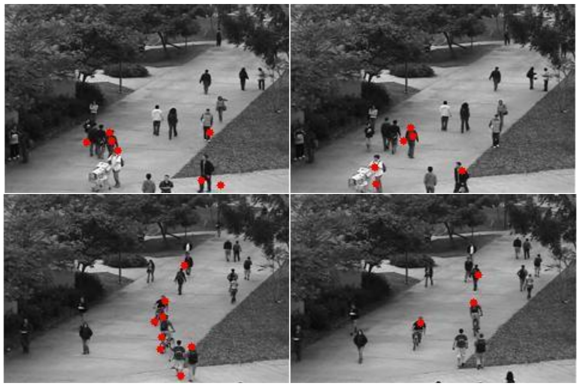


Figure 1. Examples of Spatio-temporal Interesting Points Detected by Using the Method in [8]

After the interest points are detected, in order to demonstrate the appearance and motion information, the cuboids should be extracted around the points. Specifically, cuboids have a side length of approximately six times the scale at which they are detected, as shown in Figure 2.

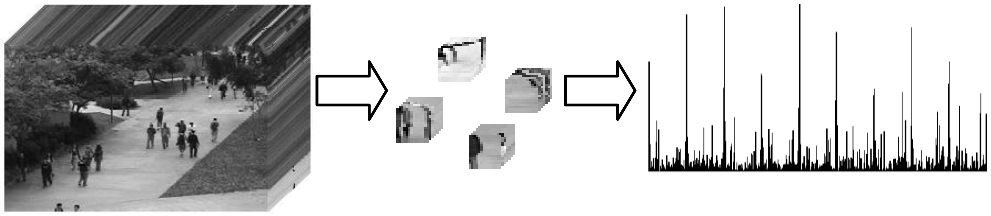


Figure 2. From Left to Right: Input Video; The Examples of Cuboids Extract-ed by Using the Method in [8]; An Example of the Representation for a Cuboid

2.2. LBP-TOP Descriptor for Spatio-Temporal Cuboids

The number of cuboids detected in videos is very large, so a simple descriptor should be defined to compute the similarity between two cuboids. Various methods can be employed to create a feature vector as a descriptor, such as flattening all the pixels into a vector, creating a histogram of optical flow or gradients. Multi-scale histogram of optical flow (MHoF) is proposed to describe each detected interest point [19]. Each detected interest point [20] is described by using histogram of gradient (HoG) and histogram of optical flow (HoF). However, these descriptors are very complicated. The designed LBP-TOP descriptor that is not only simple but discriminative can overcome the problem.

In [9], LBP-TOP descriptor for facial expressions recognition is proposed by Zhao and Pietikainen. It is developed from LBP operator [23] that describes the local texture pattern with a binary code. LBP-TOP describes DT by three orthogonal planes of a space time volume. The labels from the xy plane contain appearance information, and the labels from the xt and yt planes contain concurrent statistics of motion in horizontal and vertical directions. These three histograms are concatenated to build a global descriptor of DT with the spatial and temporal features. Binary code makes the descriptor simple with small computation. The LBP-TOP descriptor is a statistical one, which cannot describe any special information or position relationships between interesting points, similar to the limits of the Bag of Words (BoW) method. In order to overcome this shortcoming, Zhao and Pietikinen [9], by dividing the facial area into several parts and computing the LBP-TOP features of each part as local information, yield the global description for the facial expression. Figure 3 illustrates the process of computing the LBP-TOP descriptor.

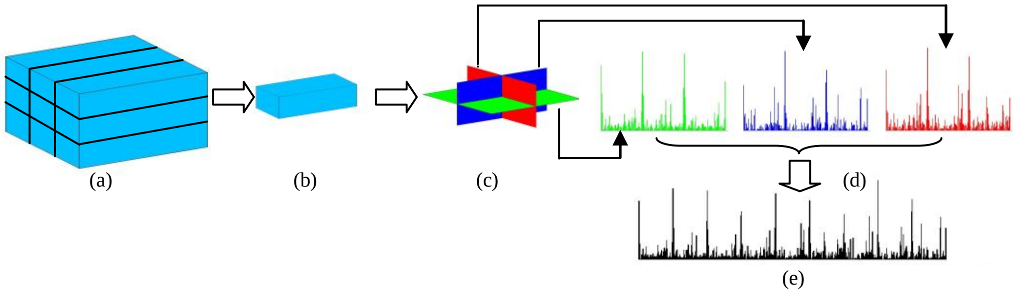


Figure 3. An Illustration of the Process of Computing the LBP-TOP for a Cuboid. (a) A Cuboid Divided into Several Sub-cuboids (9 for Example), (b) One Sub-cuboid from an Observed Cuboid, (c) Three Planes in Dynamic Texture, (d) An LBP Histogram from Each Plane (e) Forming an LBP-TOP Feature by Concatenating the Feature Histogram for a Sub-cuboid, (e) Concatenating All the LBP-TOP Features of the Sub-cuboids (9 for Example) as the Final Feature for the Observed Cuboid

Inspired by [9], the descriptor is used for human detection and human action recognition [10]. In their work, only xt and yt are considered, however, the xy plane contains a lot of useful appearance information. In general, the value of the LBP code is given by

$$LBP(x_c, y_c) = \sum_{p=0}^{p-1} s(g_p - g_c)2^p, \quad s(x)\begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

where g_c is the gray value of the center pixel (x_c, y_c, t_c), g_p is the gray value found at the p sampling points: $(x_c - R_x \sin(2\pi p / P_{xy}), y_c + R_y \cos(2\pi p / P_{xy}), t_c)$ for the xy plane, the coordinates of g_{xt} are given by $(x_c - R_x \sin(2\pi p / P_{xt}), y_c, t_c - R_t \cos(2\pi p / P_{xt}))$ for the xt plane, and similarly $(x_c, y_c - R_y \cos(2\pi p / P_{yt}), t_c - R_t \sin(2\pi p / P_{yt}))$, with the coordinates of g_{yt} for the yt plane. R_d is the radius of the ellipse in the direction of the axis $d\,(x, y\ or\ t)$. The values for g_p, for points that do not fall on pixels, are estimated by using bilinear interpolation.

Motivated by these works, the LBP-TOP descriptor is applied as the descriptor of the spatio-temporal interest cuboids in this paper. Given an interest cuboid as DT, the central part is only considered to calculate the LBP-TOP feature. A histogram of the dynamic texture is defined as:

$$H_{i,j} = \sum_{x,y,t} I\left\{ f_j(x,y,t) = i \right\}, \quad i = 0,\cdots,n_j - 1;\ j = 0,1,2. \tag{2}$$

where n_j is the number of different labels produced by the LBP operator in the jth plane ($j = 0: xy, 1: xt, 2: yt$), $f_j(x,y,t)$ expresses the LBP code of the central pixel (x,y,t) in the jth plane, where I is defined as:

$$I(A) = \begin{cases} 1, if \quad A\ is\ true \\ 0, if \quad A\ is\ false \end{cases} \tag{3}$$

The histograms of the three planes are concatenated after being normalized.

3. Abnormal Event Detection Using Sparse Coding

An observed cuboid is first sparsely coded over the trained dictionary, and then the abnormal event is detected by computing the reconstruction error that is bigger than the predefined threshold.

3.1 Sparse Coding

Sparse coding can be considered evolving from the Bag of Words (BoW) approach. In the BoW method, a similar philosophy is adopted, wherein the whole descriptor of the observed data is described by the statistics of words, which is called codewords. The distribution of a small collection of the codewords is used to categorize data. The BoW approach is formulated as equation (4):

$$fea_i = \frac{1}{M} fr_i, \ i = 1,2,..., k \tag{4}$$

The histogram $Fea = [fea_1, fea_2, ..., fea_k]'$ of an observed data is represented by unordered statistics of local descriptors, where k is the number of codewords and fr_i is the frequency of the ith codeword. The local descriptors are obtained by computing the similarity between local features and codewords. In other words, each local feature is described by only one of the codewords. For example, in human action recognition, many spatio-temporal interest points have been extracted for all actions and a dictionary is formed by using clustering methods such as K-means. The spatio-temporal interest points coming from an action video are used to characterize this video. Each spatio-temporal interest point is considered to be represented by only one codeword:

$$\underset{i}{\arg} \min_{i=1,2,...,k} \left\| x_j - w_i \right\|^2 \tag{5}$$

where x_j is the jth spatio-temporal interest point in the set. If the sample x_j is identified as the ith word, the corresponding frequency value of ith word in the video increases by one. Finally, the frequency of each codeword is computed and normalized as a feature of the whole video. It is so rough that there will be deviation from the original feature. To address this problem, the local feature can be represented by several codewords, in other words, the sample x_j can be constructed by some of words in the dictionary. The method is called sparse coding. Equation (5) is updated in (6):

$$\underset{\alpha}{\arg} \min \left\| x_j - W\alpha \right\|_2^2 \tag{6}$$

where $W = [w_1, w_2, ..., w_k]$ is the dictionary. To ensure that α has a small number of nonzero elements, equation (6) is evolved into:

$$\underset{\alpha}{\arg} \min \left\| x_j - W\alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1 \tag{7}$$

where$W \in \Gamma = \{W : w_j^T w_j \leq 1, \forall j = 1,2,\cdots,k\}$, and$\lambda$ is the sparsity enforcer. The representation has a higher accuracy than BoW. W is an underdetermined matrix in sparse coding, therefore there will be many solutions. However, L_1 norm of α ensures that there will be only one solution for the restrictions in equation (7). The optimization problem is convex in the dictionary W with α is fixed, and convex in coefficient α with W fixed. However, it is not jointly convex in both W and α simultaneously. The conventional way to solve the problem is to alternate between these two variables, minimizing one while clamping the other constant. The problem is solved by using the algorithm proposed by [15].

3.2 Abnormal Event Detection

The typical steps in the framework are: (1) spatio-tempral cuboids and the corresponding descriptors are extracted from all the images in the training set, (2) an over-completely dictionary is trained by using sparse coding, (3) each cuboid in the testing dataset is represented by the bases of the dictionary and (4) a cuboid is detected as an abnormal event by computing the residual, which is bigger than the threshold predefined by the user.

Given a set of training videos where only normal events occur, the dictionary can be obtained by solving equation (7) using the algorithm proposed by [15]. If x_j is a normal event, it should be represented by a small number of codewords in the dictionary. This means that the residual in (8) is very small. On the other hand, if x_j is an abnormal event, it is not represented by a small number of codewords in the dictionary, which means that residual J is very large, even if it is represented by the unusual event from all the codewords in the dictionary:

$$J = \left\| x_j - W\alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1 \tag{8}$$

In the work, Orthogonal Matching Pursuit (OMP) is used to obtain the coefficient α after the dictionary is learned. Intuitively, x_j is detected as an unusual event by finding the threshold of J . If the following criterion is satisfied, then x_j is an unusual event:

$$J(x_j, \alpha, W) \geq \varepsilon \tag{9}$$

where ε is the user predefined threshold, which can control the sensitivity to the abnormal event. Figure 4 provides the reconstruction code of a spatio-temporal cuboid. We restrict the number of several significant coefficients in OMP algorithm. With the limit of the number of nonzero elements, the reconstruction errors of abnormal cuboids are bigger than the threshold.

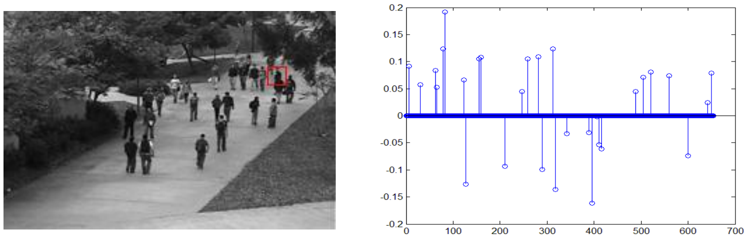


Figure 4. Example Frame and Sparse Code

4. Experiments and Comparisons

We apply the published UMN dataset [24] and UCSD dataset [25] to test the effectiveness of our proposed algorithm. The UMN dataset is used to test the global abnormal event detection; and the UCSD dataset is used to detect the local event detection.

4.1 UCSD Dataset and Experimental Results Analysis

The UCSD Ped1 Dataset consists of 34 normal event videos and 36 abnormal event videos. Each video has 200 frames, with a resolution of 238 × 158, and 10 fps. The normal event videos contain pedestrians in walkways. The abnormal event videos contain bikers, skaters, cars and pedestrians walking in anomalous motion patterns, or in non-walkway regions. The typical procedure is: (1) extract the spatio-temporal interest cuboids in all the images in the UCSD dataset, where the size of cuboids extracted is $13 \times 13 \times 31$ in our experiments, (2) obtain corresponding LBP-TOP descriptors, being a 177×1 vector, (3) train the dictionary using equation (7) in a normal training set. The dictionary size is 654 and the non-zero element number in coefficient α is 30. (4) The coefficient α is computed with the fixed dictionary and the residual is compared with the threshold ε . Extract 11,077 spatio-temporal interest cuboids in the training set and 21,141 cuboids in the test set, with the parameters $\sigma = 1$ and $\tau = 2.5$ for the Gaussian and Gabor filters. In addition, the threshold is 0.02 for the response of the filters. The cuboids are extracted, where the maximal value of responses are larger than this threshold. The threshold for abnormal event detection is 0.1.

The examples of detection results are shown in Figure 5. The algorithm can detect bikers, skaters, small cars, incorrect direction and walking in lawn. In Figure 6, the proposed method is compared with MDT, Social force and MPPCA, and other methods. It is easy to see that the ROC curve out-performs the others. In Table 1, the Equal Error Rate (EER) and Area under Curve (AUC) evaluation results are presented. The EER for the proposed method is 12.4%, which is less than the closest EER of 19% [19], and the AUC is 93.61%, which is greater than the closest AUC of 90.99% [19], thus it can be concluded that the performance of the proposed algorithm outperforms the state-of-the-art methods.

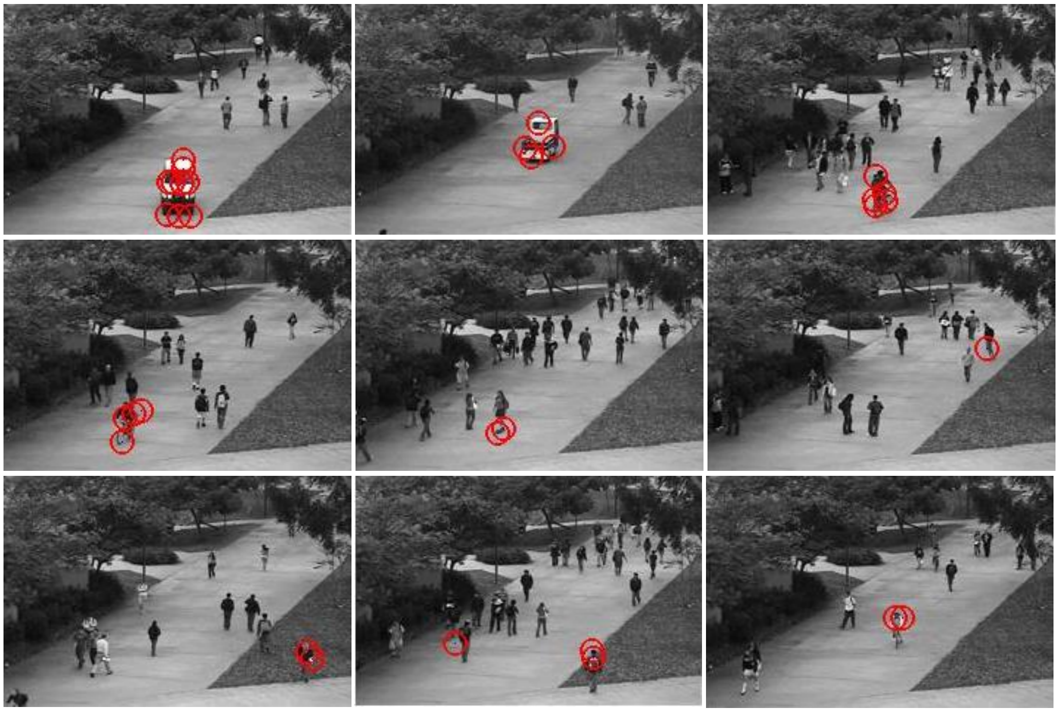
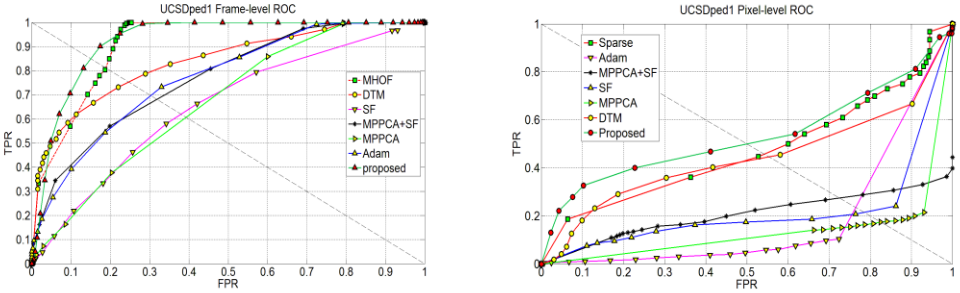


Figure 5. Examples of Abnormal Event Detection for UCSD Ped1 Datasets (The Bikers, Skaters, Vehicles, Wrong Direction and Walking in the Lawn)



(a)Frame-level ROCs for the Ped1 Dataset (b) Pixel-level ROCs for Ped1 Dataset

Figure 6. The Detection Results of the UCSD Ped1 Dataset

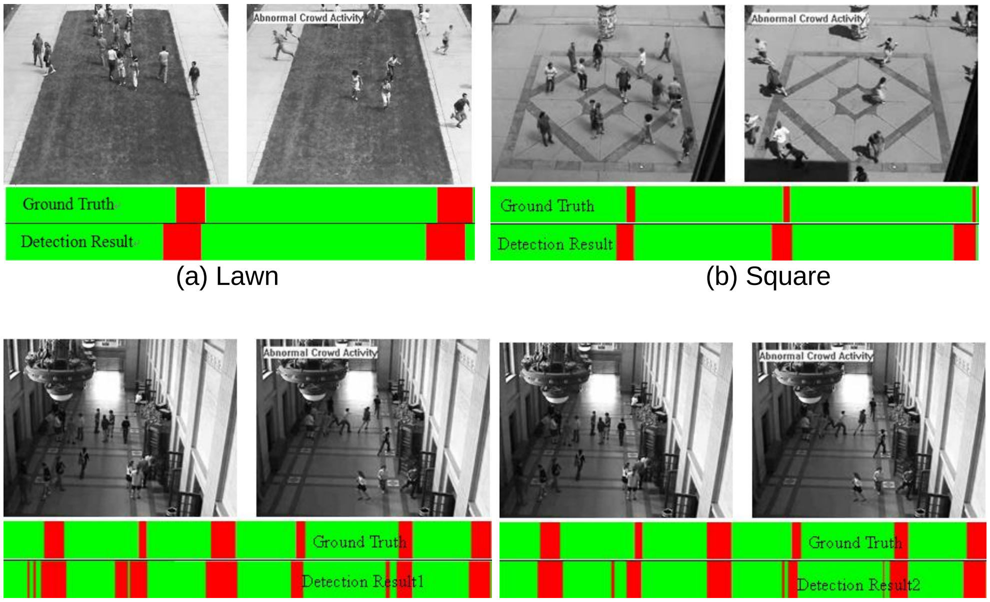
Table 1. The Comparison of EER and AUC on the Ucsdped1 Dataset

	EER	*AUC*
Adam[15]	38%	77.05%
SF[16]	31%	58.24%
MPPCA[16]	40%	67.03%
MPPCA+SF[16]	32%	76.96%
DTM[16]	25%	83.77%
MHOF+Sparse[9]	19%	90.99%
The proposed	15%	92.43%

4.1 UMN Dataset and Experimental Results Analysis

For the UCSD dataset, the abnormal event detection is to detect local interest events, which means that there exist normal and abnormal events at the same time in the whole scenario. Different from UCSD dataset, there exist only normal events or abnormal events at the same time in the videos of the UMN dataset, which is called global abnormal event detection. The UMN dataset consists of 3 different scenes of crowded escape events, and the total frame number is 7740(1450, 4415 and 2145 for scenes 1−3, respectively) with a 320×240 resolution. The trained dictionary is initialized with the first 400 frames of scene 1 and scene 3, the normal 400 frames of scene 2 for the presence of abnormal events in the first 400 frames. And the left frames are used for testing. There are some special cases for scene 2. It is critical whether the frames used to train the dictionary include lighting changes. The trained dictionary size is 300 and the non-zero element number in coefficient α is 10. The threshold defined for abnormal event detection is 0.185. The threshold for extracting interest cuboids is 0.005, with the parameters $\sigma = 1.5$ and $\tau = 2.5$ for the Gaussian and Gabor filters.

The results of the experiments are shown in Figure 7. The overall results show that the proposed approach is capable of detecting each abnormal event annotated in the database, while maintaining a low number of false positives. For scene 2(hall), the detection result 1 is the outcome for the dictionary trained without using frames of light changes, and detection result 2 is obtained using the dictionary trained with frames of light changes. It is clear that detection result 2 shows better performance.

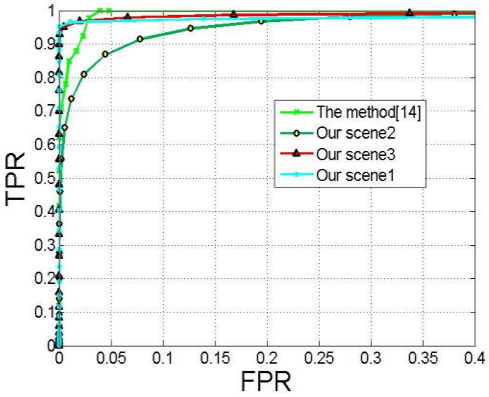


(a) Lawn (b) Square

(c) Hall (d) Hall

Figure 7. The Qualitative Results of the Abnormal Event Detection for the Three Scenarios of the UMN Dataset, the Ground Truth and the Detection Result Bars Show the Labels of Each Frame, Where Green Color Denotes the Normal Frames and Red Corresponds to Abnormal Frames

In Figure 8, the ROC curves, by frame-level measurement, are shown in order to compare to the state-of-the-art methods, and the quantitative comparisons to the other three methods are also provided. Generally, the detections of the proposed systems are longer, and begin at an earlier time, than the annotated ground truth. Some frame images have the abnormal activity labeled very late, after the person started to run, which explains the delay as the proposed system reacts immediately to motion changes.



	Area under ROC
Chaotic Invariants [26]	99%
Social Force [27]	96%
Optical flow [27]	84%
Scene 1 [14]	99.5%
Scene 2 [14]	97.5%
Scene 3 [14]	96.4%
Our Scene1	98.2%
Ours Scene2	97.5%
Ours Scene3	99.2%

(a) Frame-level ROCs for UMN Dataset (b) Quantitative Comparison of Our Method with [14], [26] and [27]

Figure 8. The Detection Results of the UMN Dataset

All experiments are run on a computer with 2GB RAM and a 2.4GHz CPU. The average computation time is 5.52 sec/frames for UMN dataset, 1.86 sec/frames for UCSD dataset. The computation time is related to the size of dictionary and the number of interest points.

5. Conclusion

The method of combining LBP-TOP descriptor with the sparse coding is proposed for abnormal event detection in crowded scenarios. The LBP-TOP descriptor based on the binary code is more discriminative and has lower computation cost than other descriptors. Therefore, it is appropriate for abnormal event detection with thousands of cuboids. Different from the traditional sparse coding, in the testing phase, the OMP method is used to compute the coefficients for the test cuboids. Experiments on the UCSD and UMN databases, with a comparison to state-of-the-art results, demonstrate that the proposed method obtains good performance. The AUC for the UCSD dataset is 93.61%, and the UMN is 98.2%, 97.5% and 99.2%. In the UMN dataset, scene 2 has large light changes, which affects the results. Moreover, only 400 frames are used to learn the dictionary. In the future, the accuracy could be further improved by learning the dictionary using more frames or using multi-features.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant no.61103123.

References

[1] V. Chandola, A. Banerjee and V. Kumar, “Anomaly Detection: A survey”, ACM Comput, Surveys, vol. 41, no. 3, **(2009)**, pp. 1-58.

[2] D. Anderson, R. Luke, J. Keller, M. Skubic, M. Rantz and M. Aud, “Linguistic Summarization of Video for Fall Detection Using Voxel Person and Fuzzy Logic”, Computer Vision and Image Understanding, vol. 113, no. 1, (**2009**), pp. 80-89.

[3] O. Boiman and M. Irani, “Detecting Irregularities in Images and In Video”, IEEE International Conference on

Computer Vision, (**2005**), pp. 462-469.

[4] A. Nasution and S. Emmanuel, "Intelligent Video Surveillance for Monitoring Elderly in Home Environments", IEEE Workshop on Multimedia Signal Processing, (**2007**), pp. 203-206.

[5] F. Nater, H. Grabner and L. Van Gool, "Exploiting Simple Hierarchies for Unsupervised Human Behavior Analysis", IEEE Conference on Computer Vision and Pattern Recognition, (**2010**) pp. 2014-2021.

[6] F. Nater, H. Grabner, T. Jaeggli and L. Van Gool, "Tracker Trees for Unusual Event Detection", IEEE Workshop on Visual Surveillance, (**2009**).

[7] A. Wiliem, V. K. Madasu, W. W. Boles, P. K. Yarlagadda, "Adaptive Unsupervised Learning of Human Actions", The 3rd International Conference on Imaging for Crime Detection and Prevention, (**2009**).

[8] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, "Behavior Recognition via Sparse Spatio-temporal Features", IEEE International Workshop On Visual Surveillance and Performance Evaluation of Tracking and Surveillance, (**2005**), pp. 65-72.

[9] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns With an Application to Facial Expressions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, (**2007**), pp. 915 – 928.

[10] V. Kellokumpu, G. Zhao and M. Pietikäinen, "Human Activity Recognition Using a Dynamic Texture Based Method", British Machine Vision Conference, (**2008**), pp. 1-10.

[11] A. Y. Y. Wright, A. Ganesh, S. S. Sastry and Y. Ma, "Robust Face Recognition via Sparse Representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, (**2009**), pp. 210-227.

[12] M. Yang, L. Zhang, J. Yang and D. Zhang, "Robust Sparse Coding for Face Recognition", IEEE Conference on Computer Vision and Pattern Recognition, (**2011**), pp. 625-632.

[13] L. Zhang, M. Yang and X. Feng, "Sparse Representation or Collaborative Representation: Which Helps Face Recognition", IEEE International Conference on Computer Vision, (**2011**), pp. 471-478.

[14] J. Yang, K. Yu, Y. Gong and T. Huang, "Linear Spatial Pyramid Matching using Sparse Coding for Image Classification", IEEE Conference on Computer Vision and Pattern Recognition, (**2009**), pp. 1794-1801.

[15] H. Lee, A. Battle, R. Raina and A. Y. Ng, "Efficient Sparse Coding Algorithms", Advances in Neural Information Processing Systems, (**2006**), pp. 801-808.

[16] Q. Qiu, Z. Jiang and R. Chellappa", Sparse Dictionary-based Representation and Recognition of Action Attributes", IEEE International Conference on Computer Vision, (**2011**), pp. 707-714.

[17] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas and L. Fei-Fei, "Human Action Recognition by Learning Bases of Action Attributes and Parts", IEEE International Conference on Computer Vision, (**2011**), pp. 1331-1338.

[18] Y. Liu and Y. Li, "Human Action Recognition in Videos Using Distance Image Volumes and Sparse Coding", Journal of Computational Information System, vol. 8, no. 9, (**2012**), pp. 3557-3564.

[19] Y. Cong, J. Yuan and J. Liu, "Sparse Reconstruction Cost for Abnormal Event Detection", IEEE Conference on Computer Vision and Pattern Recognition, (**2011**), pp. 3449-3456.

[20] B. Zhao, F.-F. Li and E. P. Xing, "Online Detection of Unusual Events in Videos via Dynamic Sparse Coding", IEEE Conference on Computer Vision and Pattern Recognition, (**2011**), pp. 3313-3320.

[21] C. Zhao, X. Wang and W.-K. Cham, "Background Subtraction via Robust Dictionary Learning", EURASIP J. Image and Video Processing, (**2011**).

[22] Q. Wang, F. Chen, W. Xu and M.-H. Yang, "Online Discriminative Object Tracking with Local Sparse Representation", IEEE Workshop on the Applications of Computer Vision, (**2012**), pp. 425-432.

[23] T. Ojala, M. Pietikainen and T. Maenpaa, "Multi-resolution gray-scale and rotation invariant texture classification with local binary patterns", IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, (**2002**), pp. 971-987.

[24] "Unusual crowd activity dataset of University of Minnesota", from http://mha.cs.umn.edu/movies.

[25] V. Mahadevan, W. Li, V. Bhalodia and N. Vasconcelos, "Anomaly Detection in Crowded Scenes", IEEE Conference on Computer Vision and Pattern Recognition, **(2010)**, pp. 1975-1981.

[26] M. S. R. Mehran and A. Oyama, "Abnormal Crowd Behavior Detection Using Social Force Model", IEEE Conference on Computer Vision and Pattern Recognition, (**2009**), pp. 935-942.

[27] S. Wu, B. Moore and M. Shah, "Chaotic Invariants of Lagrangian Particle Trajectories for Anomaly Detection in Crowded Scenes", IEEE Conference on Computer Vision and Pattern Recognition, (**2010**), pp. 2054-2060.

Authors

Yang Liu, she received her M.S. degree from the Northeastern University in China in 2004. She is a doctoral student at the Nanjing University of Aeronautics and Astronautics. Her research interests include vision analysis and pattern recognition.

Yibo Li, he received his M.S. and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics and Northeastern University, in 1986 and 2003, respectively. Since 1999, he holds the position of Full Professor at Shenyang Aerospace University. He has published over 100 technical research papers and books. More than 30 research papers have been indexed by SCI/EI. His research interests include vision analysis and pattern recognition.

Xiaofei Ji, she received her M.S. and Ph.D. degrees from the Liaoning Shihua University and University of Portsmouth, in 2003 and 2010, respectively. From 2013, she holds the position of Associate Professor at Shenyang Aerospace University. She is the IEEE member, has published over 20(indexed by SCI/EI) technical research papers. Her research interests include vision analysis and pattern recognition. She is the leader of National Natural Science Fund Project (Number: 61103123).