ABNORMAL EVENT DETECTION BASED ON SPARSE RECONSTRUCTION IN CROWDED SCENES

Ang Li, Zhenjiang Miao, Yigang Cen, Qinghua Liang

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
lianghit@126.com, zjmiao@bjtu.edu.cn, ygcen@bjtu.edu.cn, liangqinghua@bjtu.edu.cn

ABSTRACT

In this paper, we propose an algorithm of abnormal event detection in crowded scenes using sparse representation over the bases of normal motion feature descriptors. To construct an over-complete dictionary, we extract the histogram of maximal optical flow projection (HMOFP) feature from a set of normal training frames. Then the K-SVD dictionary training method is used to get a redundant dictionary after a process of selecting the training samples, which is better than the dictionary simply composed by the HMOFP feature of the whole training frames. In order to detect whether a frame is normal or not, we use the l_1-norm of the sparse reconstruction coefficients (i.e., the sparse reconstruction cost, SRC) to show the anomaly of the testing frame, which is simple but very effective. The experiment results on UMN dataset and the comparison to the state-of-the-art methods show that our algorithm is promising.

Index Terms—HMOFP, K-SVD, Sparse representation, Abnormal events, Crowded scenes

1. INTRODUCTION

Nowadays, with the advancement of people's public safety awareness and the reduction of surveillance equipments' cost, more and more surveillance cameras have been used in public places, such as markets, stadiums, museums, airports, train stations, etc. The research on crowd behaviors in public scenes draws more and more attention and has become a hot topic in the field of computer vision.

The algorithms of abnormal event detection can be classified into three main categories: macroscopic modeling, microscopic modeling and crowd events detection [1], [2]. Social force model based abnormal crowd behavior detection was introduced in [3]. In [4], a model named social attribute-aware force model was proposed. In [5], a novel abnormal event detection framework based on the newly developed spatial-temporal co-occurrence Gaussian mixture models (STCOG) was presented, which required a short training period and had a fast processing speed. The method using histogram of optical flow was described in [6]. A similar pixel-based motion feature HMOFP for abnormal event detection was proposed in [7].

Unlike most existing approaches used to detect abnormal events, sparse representation has obtained more and more attentions in recent years. In [8], a model aimed at anomaly detection was described, which utilized SRC over the normal dictionary to measure the normalness of the tested frame. To get an optimized dictionary in the process of sparse representation, some methods were presented, such as online dictionary learning for sparse coding [9], non-negative matrix factorization (NMF) based on the robust Earth Mover's Distance (EMD) [10], etc.

The work presented here has focused on motion feature extraction and dictionary construction. During the process of motion feature extraction, we only select one component in a feature bin, and the works in [6], [8], [9], [10] take all the components in a feature bin as the motion feature. Also, we utilize the K-SVD dictionary training method [11] after some preprocessing, which is different to previous studies.

2. MOTION FEATURE EXTRACTION

Optical flow field is the movement on the surface of grayscale images, which reflects the movement information of two consecutive frames. Optical flow can provide the information of direction and amplitude of the moving object in a scene, which can describe the behavior of people very well. In this paper, we adopt the Horn-Schunck (HS) method to compute the optical flow of frame images and propose a novel scene descriptor, called as the Histogram of Maximal Optical Flow Projection (HMOFP).

As shown in Figure 1, the optical flow field of frame s is divided into m image blocks with overlap areas, and each block contains $B_1 \times B_2$ pixels. Then we deal with the optical flow in each block as follows. $0° - 360°$ are segmented into p bins. For an image block, the optical flow vector of each pixel must belong to a bin. Thus, each bin may contain several optical flow vectors. We project all optical flow vectors in a same bin onto the angle bisector of this bin. Then the length of the maximal projection vector is selected as the feature descriptor. For example, in Figure 2(a), there are two vectors $\overrightarrow{on_1}$ and $\overrightarrow{on_2}$ falling into the first

bin. It is easy to know that the projection of $\overrightarrow{on_2}$ onto the angle bisector of the first bin is longer than that of $\overrightarrow{on_1}$. Thus, the length of the projection vector $\overrightarrow{on_{2'}}$ is selected as the feature descriptor of the first bin. After computing m blocks, we obtain the feature descriptor vector of each image block, which can be denoted as $[h_1, h_2, ..., h_m]_{p\times m}$, where $h_i = [h_i^1, h_i^2, ..., h_i^p]_{p\times 1}^T$. For the i^{th} block, $h_i^j (1 \le j \le p, 1 \le i \le m)$ denotes the maximal amplitude of all projection vectors in the j^{th} bin. As shown in Figure 2(b), we take the concatenation of the m feature descriptor vectors, which is named H_s, as the global HMOFP feature of frame s.

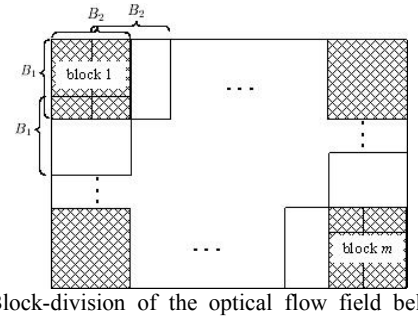
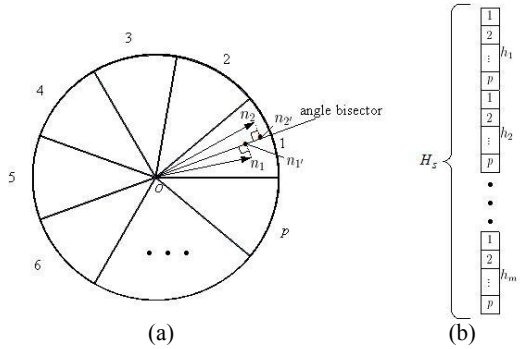


Figure 1: Block-division of the optical flow field belonging to frame s.



(a) (b)

Figure 2: (a) The calculation of the histogram of maximal optical flow projection (HMOFP) in each bin. (b) Components of the global feature descriptor of frame s.

In order to describe a crowded scene well, enough motion information of the crowd should be achieved. To describe the motion of a crowd, we need two factors: explicit directions and the moving distance along each direction. The operation of segmenting the $2D$ space into p bins provides us ample information to describe the directions of moving people. To let the direction in each bin be unique, we select the p angle bisectors as the direction standard. Since there may be far more than one optical flow vector in each bin, in order to enhance the distinction between the normal scene and the abnormal scene, we select the maximal vector projection rather than the sum of all the vector projections on the bisector as the motion feature descriptor. If we ignore the background area, the amplitudes of motion vectors belong to the normal area are very small in a normal frame and the motion vectors corresponding to the abnormal area are large in an abnormal frame. Usually, the number of normal motion vectors is much more than that of the abnormal area. If we use the sum of all projection vectors on the angle bisector as the feature descriptor of each bin, the accumulation of the massive small motion vectors in the normal frame may confuse the small number of large motion vectors in the abnormal frame, i.e., the sum of all projection vectors on the angle bisector in each bin of the normal frame is likely to be close to that of the abnormal frame. Thus, in order to improve the distinguishability between the abnormal and normal frames, we select the length of the maximal projection vector as the feature descriptor of each bin, as it is demonstrated in Figure 2.

3. DICTIONARY CONSTRUCTION

In this section, we address the problem how to construct the dictionary. Given an initial training set denoted as $F = [f_1, f_2, ..., f_N]$, where N is the number of frames in the set. $f_i (1 \le i \le N)$ denotes a frame image of the set and it is called a *training frame* in this paper. The corresponding feature pool is $\mathcal{H} = [H_1, H_2, ..., H_{N_0}] \in \mathbb{R}^{p\times m\times N_0}$, where $N_0 = N - 1$. $H_i (1 \le i \le N_0)$ denotes the motion feature of a training frame, and it is called a *training sample* in this paper. Our method to compute the motion feature is on the basis of optical flow, and the way to calculate optical flow based on two consecutive frames is only effective to the first frame, so the maximal subscript of H_i is $N - 1$. We realize that in the feature pool $\mathcal{H}$, there may be such training samples that have little relationship with the others. So we should do effort to delete such samples. Considering the optimization problem:

$$\min_{s_j\in\mathbb{R}^{N_0}} \|s_j\|_0 \ \text{ s.t. } \ \mathcal{H}s_j = H_j, s_{jj} = 0 \ (j = 1, 2, ..., N_0) \quad (1)$$

where $s_j = [s_{j1}, s_{j2}, ..., s_{jN_0}]^T$. (1) is the general NP-hard problem. We can use the method in [12] to relax the l_0-norm optimization problem as:

$$\min_{s_j\in\mathbb{R}^{N_0}} \|s_j\|_1 \ \text{ s.t. } \ \mathcal{H}s_j = H_j, s_{jj} = 0 \ (j = 1, 2, ..., N_0) \quad (2)$$

In the matrix form, the problem can be described as

$$\min_{S\in\mathbb{R}^{N_0\times N_0}} \|S\|_1 \ \text{ s.t. } \ \mathcal{H}S = \mathcal{H}, diag(S) = \mathbf{0} \quad (3)$$

where $S = [s_1, s_2, ..., s_{N_0}]$.

We utilize the orthogonal matching pursuit (OMP) method [13] to solve (3). After the optimal S^* is achieved, we inspect each row in it via the equation:

$$S^* = [s_T^{1\,*}, s_T^{2\,*}, ..., s_T^{N_0\,*}]^T \quad (4)$$

where $s_T^{j'\,*}, (j' = 1, 2, ..., N_0)$ is the j'^{th} row of S^*. We calculate the l_2-norm of each $s_T^{j'\,*}$. If the result is 0, we delete the corresponding column in $\mathcal{H}$. The optimized $\mathcal{H}$ is denoted as $\mathcal{H}^* = [H_1^*, H_2^*, ..., H_{K_0}^*] \in \mathbb{R}^{p\times m\times K_0} (K_0 < N_0)$. After the training sample set is optimized, the K-SVD algorithm is utilized to generate an optimal dictionary with

proper redundancy, such that the atoms in the dictionary can be more representative for the normal features. The K-SVD algorithm is described as follows.

Algorithm 1: The K-SVD algorithm

Task: Find the best dictionary to represent the training samples in $\mathcal{H}^*$ as sparse compositions, by solving

$$\min_{D,X} \|\mathcal{H}^* - DX\|_F^2 \ \text{ s.t. } \ \forall k, \|x_k\|_0 \le T_0$$

where T_0 is a small number, and $1 \le k \le K_0$.

Initialization: Set the dictionary matrix $D^{(0)} \in \mathbb{R}^{p\times m\times K}$ with l_2 normalized columns. Set $J = 1$.

Repeat until convergence (stopping rule):

(1) Sparse Coding Stage: Use the OMP algorithm to compute the representation vectors x_k for each example H_k^*, by approximating the solution of

$$\min_{x_k} \|H_k^* - Dx_k\|_2^2 \ \text{ s.t. } \ \|x_k\|_0 \le T_0$$

(2) Code book Update Stage: For each column in $D^{(J-1)}$, update it by

(a) Define the group of examples that use this atom,

$$w_{k'} = \{k | 1 \le k \le K_0, x_T^{k'}(k) \ne 0\} (k' = 1, 2, \ldots, K)$$

(b) Compute the overall representation error matrix, $E_{k'}$, by

$$E_{k'} = \mathcal{H}^* - \sum_{t \ne k'} d_t x_T^t (1 \le t \le K).$$

(c) Restrict $E_{k'}$ got from (b) by choosing only the columns corresponding to $w_{k'}$, and obtain $E_{k'}^{\mathcal{R}} = E_{k'}\Omega_{k'}$, where $\Omega_{k'}$ is defined as a matrix of size $K_0 \times |w_{k'}|$, with ones on the $(w_{k'}(k), k)^{th}$ entries and zeros on the other entries.

(d) Apply SVD decomposition $E_{k'}^{\mathcal{R}} = U\Delta V^T$. Choose the updated dictionary column $\widetilde{d}_{k'}$ to be the first column of U. Update the coefficient vector $x_{\mathcal{R}}^{k'} = x_T^{k'}\Omega_{k'}$ to be the first column of V multiplied by $\Delta(1,1)$.

(3) Set $J = J + 1$.

When this process of the K-SVD algorithm ends, we obtain an optimized redundant dictionary.

4. ABNORMAL EVENT DETECTION

In this section, an algorithm to detect abnormal events in surveillance video is described in detail. Suppose that in a given scene, there is a set of training frames, $F = [f_1, f_2, ..., f_N]$, which describe the normal behavior of crowded people. The general procedures to detect the newly incoming frames based on the histogram of maximal optical flow projection (HMOFP) feature are introduced as follows.

Step 1: Calculate the optical flow of the training frames, i.e., $OP = [op_1, op_2, ..., op_{N_0}]$, at each pixel of the first N_0 frames by the HS method:

$$[f_1, f_2, ..., f_N]_{a\times b\times N} \xrightarrow{\text{HS}} [op_1, op_2, ..., op_{N_0}]_{a\times b\times N_0} \quad (5)$$

where $a \times b$ is the size of the frame image in the initial training set.

Step 2: Extract the motion feature HMOFP of the first N_0 training frames in the training set, which is described as the set $[H_1, H_2, ..., H_{N_0}]$.

$$[op_1, op_2, ..., op_{N_0}]_{a\times b\times N_0} \xrightarrow{\text{HMOFP}} [H_1, H_2, ..., H_{N_0}]_{p\times m\times N_0} \quad (6)$$

Step 3: Based on HMOFP, we delete the useless columns in $\mathcal{H}$ and get the optimized dictionary D with the K-SVD algorithm as introduced in section 3.

Step 4: Get the HMOFP feature of the incoming frame f_{t_0} and calculate the l_1-norm of the sparse reconstruction coefficient vector z_{t_0} with OMP method over the dictionary D, which is denoted as

$$S_w = \|z_{t_0}\|_1 \quad (7)$$

where S_w is the SRC value. The frame f_{t_0} is detected as normal if the following criterion is satisfied $S_w < \tau$, where τ is a user defined threshold that controls the sensitivity of the algorithm to abnormal events.

5. EXPERIMENTAL RESULTS

There are three different crowded scenes in UMN dataset [14], which are named lawn, indoor and plaza respectively, and the total frame number is 7739 with a 320×240 resolution. The normal events are people walking randomly in the scene, and the abnormal events are human running away at the same time. In our experiments, the image block size is set as 80×60 and there is no overlapping proportion in two neighboring blocks. $0° - 360°$ are divided into 18 bins, i.e., $p = 18$. The length of the HMOFP feature is 288. The initial dictionary is a discrete cosine transform (DCT) matrix with a 288×576 size, and it is trained with the first 400 normal frames in each scene.

5.1. Detection in the lawn scene

The video sequence of the lawn scene contains 1453 frames in total (i.e., the first 1452 frames are detected). Two different events in the lawn scene are shown in Figure 3. The detection result of the lawn scene is shown in Figure 4.

5.2. Detection in the indoor scene

The video sequence of the indoor scene contains 4144 frames in total (i.e., the first 4143 frames are detected). Two different events in the indoor scene are shown in Figure 5. The detection result of the indoor scene is shown in Figure 6.

5.3. Detection in the plaza scene

The video sequence of the plaza scene contains 2142 frames in total (i.e., the first 2141 frames are detected). Two different events in the plaza scene are shown in Figure 7. The detection result of the plaza scene is shown in Figure 8.

(a) The normal event (b) The abnormal event

Figure 3: Two different events in the lawn scene.

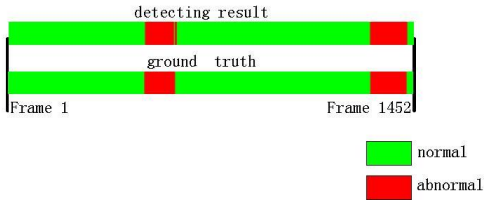


Figure 4: The classification result of the lawn scene.

(a) The normal event (b) The abnormal event

Figure 5: Two different events in the indoor scene.

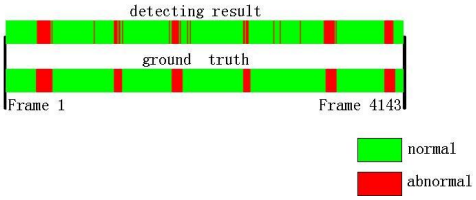


Figure 6: The classification result of the indoor scene.

(a) The normal event (b) The abnormal event

Figure 7: Two different events in the plaza scene.

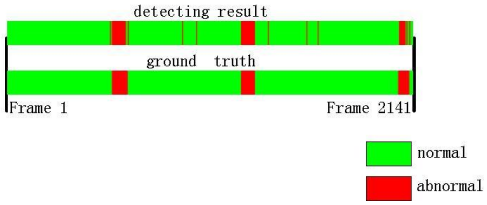


Figure 8: The classification result of the plaza scene.

5.4. The receiver operating characteristic (ROC) curve

In each scene, the ROC curve is shown in Figure 9. The area under the ROC curve (AUC) is 0.9976 in the lawn scene, 0.9570 in the indoor scene and 0.9869 in the plaza scene.

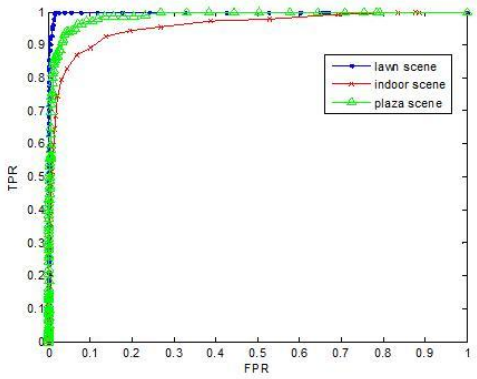


Figure 9: The ROC curves of the three scenes.

The performances of our algorithm based on the HMOFP feature and of the state-of-the-art methods are shown in Table 1. Our algorithm outperforms the methods of Optical Flow [3], NN [8], STCOG [5] and HOFO [6] and is comparable to the other methods. However, our algorithm is with a simple model and a simplified SRC form.

Method \ Scene / AUC	lawn	indoor	plaza
Social Force [3]	0.96		
Optical Flow [3]	0.84		
NN[8]	0.93		
STCOG [5]	0.9362	0.7759	0.9661
MHOF [8]	0.995	0.975	0.964
HOFO [6]	0.9845	0.9037	0.9815
Ours	0.9976	0.9570	0.9869

Table 1: The comparison of our proposed algorithm with the state-of-the-art methods.

6. CONCLUSIONS

In this work, we proposed an algorithm to detect abnormal events in crowded scenes with global-frame scale. Our method contains two main procedures: one is to compute the histogram of maximal optical flow projection (HMOFP) descriptor of the input video sequence, the other is to utilize the optimized dictionary to calculate the SRC values of testing sets. The proposed method has been tested on UMN dataset with satisfying results about abnormal event detection.

7. ACKNOWLEDGMENT

This work is supported by the NSFC (nos. 61273274, 61272028, 61572067 and 61370127), 973 Program (no. 2011CB302203), National Key Technology R&D Program of China (nos. 2012BAH01F03, NSFB4123104, FRFCU 2014JBZ004, and Z13111000191343), and Tsinghua-Tencent Joint Lab for IIT.

8. REFERENCES

[1] M. Thida, Y.L. Yong, P. Climent-Pérez, H-l. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," *Intelligent Multimedia Surveillance*, pp. 17-36, 2013.

[2] Zhan, Beibei, et al, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345-357, 2008.

[3] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 935-942, 2009.

[4] Y. Zhang, L. Qin, H. Yao, and Q. Huang, "Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 7, pp.1231 - 1245, 2015.

[5] Y. Shi, Y. Gao, R. Wang, "Real-time abnormal event detection in complicated scenes," *IEEE International Conference on Pattern Recognition (ICPR)*, pp. 3653–3656, 2010.

[6] T. Wang, H. Snoussi, "Detection of abnormal visual events via global optical flow orientation histogram," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 988-998, 2014.

[7] Ang Li, Zhenjiang Miao, Yigang Cen, Tian Wang, and Viacheslav Voronin, "Histogram of Maximal Optical Flow Projection for Abnormal Events Detection in Crowded Scenes," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 406941, 11 pages, 2015. doi:10.1155/2015/406941

[8] Y. Cong, J. Yuan, J. Liu, "Sparse reconstruction cost for abnormal event detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3449-3456, 2011.

[9] Duan, Shishi, Xiangyang Wang, Xiaoqing Yu, "A new method of abnormal event detection based on sparse reconstruction," *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, pp. 390-395, 2014.

[10] X. Zhu, J. Liu, J. Wang, C. Li, H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognition*, vol. 47, no.5, pp. 1791-1799, 2014.

[11] M. Aharon, M. Elad, A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 54, no. 11, pp. 4311-4322, 2006.

[12] D. L. Donoho, Y. Tsaic, "Extensions of compressed sensing," *Signal Processing*, vol. 86, no. 3, pp. 533-548, 2006.

[13] J. Tropp, A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655-4666, 2007.

[14] UMN, Unusual crowd activity dataset of University of Minnesota, department of computer science and engineering. *http://mha.cs.umn.edu/movies/crowd-activity-all.avi*, 2006.