Abnormal Activity Detection Using Spatio-Temporal Feature and Laplacian Sparse Representation

Yu Zhao[1], Yu Qiao[1,2], Jie Yang[1,2(✉)], and Nikola Kasabov[3]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
{zhaoyunccq,qiaoyu,jieyang}@sjtu.edu.cn
[2] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, China
[3] Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand
nkasabov@aut.ac.nz

Abstract. Abnormal activity detection in a video is a challenging and attractive task. In this paper, an approach using spatio-temporal feature and Laplacian sparse representation is proposed to tackle this problem. To detect the abnormal activity, we first detect interest points of a query video in the spatio-temporal domain. Then normalized combinational vectors, named HNF, are computed around the detected space-time interest points to characterize the video. After that, we utilize the Laplacian sparse representation framework and maximum pooling method to gain a more discriminative feature vector from the HNF set. Finally, the support vector machine (SVM) is adopted to classify the feature vector as normal or abnormal. Experiments on two datasets demonstrate the satisfactory performance of the proposed approach.

Keywords: Abnormal activity detection · Space-time interest points · Laplacian sparse representation · Maximum pooling · SVM

1 Introduction

In surveillance videos, the abnormal activities can be defined as aberrant events such as people fights, crowded escape activities. Thus, the abnormal activity detection is to identify these aberrant events from normal ones, which is a two-class classification problem. This problem is divided into the detection of global abnormal activity (GAA) and local abnormal activity (LAA). For GAA, there is only normal activity or abnormal activity at the same time in the entire scenario. For LAA, normal and abnormal activities emerge simultaneously in the scenario.

In some existing surveys, the global features are used to describe video clips for GAA. In [8], the streakline representation based on Lagrangian framework for fluid dynamics is utilized to detect abnormal activity. In [9], Mehran et al.

S. Arik et al. (Eds.): ICONIP 2015, Part IV, LNCS 9492, pp. 410–418, 2015.
DOI: 10.1007/978-3-319-26561-2_49

use the social force model [4] to analyze the human activity. The description of human activity in social force model is based on the intention of movement.

On the other hand, some surveys [10,11] utilize the Bag-of-Words (BoW) framework to tackle both GAA and LAA problem. The BoW model consists of three modules: (1) feature extraction; (2) codebook producing and feature quantization; (3) classification.

For the feature extraction, [10,11] use the local spatio-temporal feature to characterize video clips. They first detect the space-time interest points (STIP) [5] of video clips. Then spatio-temporal descriptors such as Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) are computed respectively around the detected STIP to describe video clips. However, HOG or HOF is not discriminative enough to encode the information of appearance and action. In [6], the HOG and HOF descriptors are combined into normalized vectors for the recognition of human actions. The combinational feature vectors of HOG and HOF, named HNF, are discriminative enough to describe the appearance and action in video clips. Note that detecting abnormal activity also needs to encode the appearance and action information. Therefore, the HNF is also a discriminative descriptor for the abnormal activity detection.

For the second module, BoW model first utilizes the K-means clustering to produce the codebook which contains several visual words. Then each feature is only assigned to its nearest visual word to generate a frequency histogram of visual words, which is called the feature quantization. However, the information loss [1] in feature quantization is severe because of the hard assignment approach used in BoW model. In order to reduce the information loss, [12] adopts the general sparse coding to generate the codebook and quantize features in abnormal activity detection, which achieves better performance than BoW model.

However, the general sparse coding disposes features separately, ignoring the similarity among features, which decreases the accuracy and robustness of sparse representation. In order to tackle this problem, Gao et al. [3] propose the Laplacian Sparse Representation (LSR) approach for image classification. In the LSR approach, the similarity matrix is used to preserve the similarity information among features, which can further reduce the feature quantization error and make the sparse representation more robust. Thus, we can also utilize the LSR

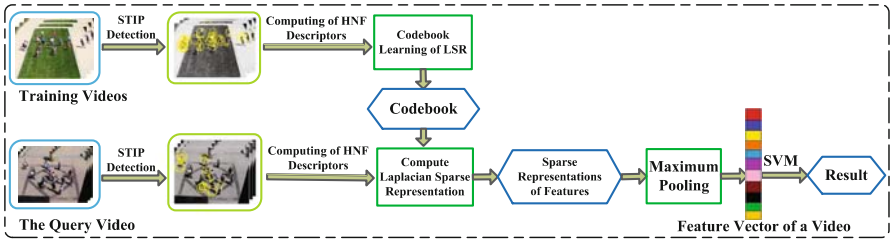

Fig. 1. The main framework of our approach

approach to learn a better codebook and produce a more descriptive video representation for abnormal activity detection.

As mentioned above, we propose an approach using the spatio-temporal feature and the Laplacian sparse representation for both GAA and LAA problem. The main framework is summarized in Fig. 1. Firstly, we detect the space-time interest points of a query video. Then the HNF descriptors are computed in the nearby 3D patches of the detected interest points. After that, the sparse representations of HNF descriptors are generated by applying the Laplacian sparse coding. Then we utilize the maximum pooling method among the entire Laplacian sparse representation set of the query video to obtain a more descriptive feature vector. Finally, we use the SVM to classify this feature vector and determine whether the query video includes the abnormal activity.

The rest of this paper is organized as follows: we present the details of our approach in Sect. 2. In Sect. 3, we report the experimental results and show the comparisons of different methods. We give the conclusion in Sect. 4.

2 Methodology

2.1 Spatio-Temporal Feature

In order to detect the abnormal activity, we first extract the spatio-temporal features from a query video, which contains two procedures. The Fig. 2 illustrates the extraction of spatio-temporal features.

Detection of Space-Time Interest Points. Detecting interest points in a video clip is to locate the local regions with intense variation of image intensity in the spatio-temporal domain. In this paper, we utilize the approach proposed by [5] to solve this problem.

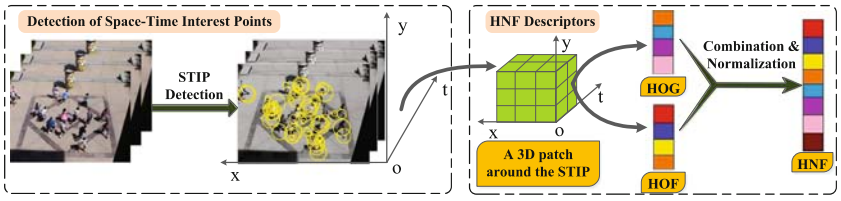

Fig. 2. The illustration of extracting spatio-temporal features

Firstly, let $f_{st}\colon \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ represent a video clip. Then, we convert the f_{st} into linear scale space by applying the convolution of f_{st} with an anisotropic Gaussian kernel. After that, we compute the second moment matrix with respect to the linear scale-space representation of f_{st}. Then we integrate the second moment matrix with an Gaussian kernel to obtain a matrix Γ_{st}. Eigenvalues

with large values of the matrix Γ_{st} indicate the local bricks which have intense variation of image intensity. Thus, following the work [5], the extended Harris corner function H_{st} with respect to the eigenvalues of Γ_{st} is introduced to tackle the problem.

$$H_{st} = \lambda_1\lambda_2\lambda_3 - \xi \cdot (\lambda_1 + \lambda_2 + \lambda_3)^3 = \lambda_1^3(\omega\eta - \xi \cdot (1 + \omega + \eta)^3) \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3$ denote the eigenvalues of matrix Γ_{st}, $\omega = \lambda_2/\lambda_1$, $\eta = \lambda_3/\lambda_1$, $\xi \leq \omega\eta/(1 + \omega + \eta)^3$, and $H_{st} \geq 0$. We can detect space-time interest points by finding the local positive maximum values of H_{st}.

HNF Descriptors for STIP. After the detection of interest points, we use the HNF descriptors [6] to characterize the video. Firstly, Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF) are computed respectively in the nearby 3D patches of detected interest points. Then, to obtain the HNF descriptors, we combine the HOG and HOF into vectors and normalize them. As shown in Fig. 2, the 3D video patch is divided into a grid with $3 \times 3 \times 2$ spatio-temporal bricks. In each brick, there are 4-bin HOG descriptors and 5-bin HOF descriptors. Therefore, one HNF descriptor vector has 162 dimensions, including 72 elements from HOG and 90 elements from HOF. With the combination of HOG and HOF, the HNF descriptor can provide more information of appearance and action.

2.2 Laplacian Sparse Representation for Spatio-Temporal Feature

In order to produce a more descriptive and precise representation for the appearance and action of a video clip, we utilize the Laplacian Sparse Representation (LSR) method [3] to encode spatio-temporal feature vectors. In the LSR method, the matrix S is used to preserve the similarity information among features and a regularization term with respect to the similarity matrix S is applied to improve the robustness and accuracy in feature quantization.

Let $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{d \times N}$ denote the spatio-temporal features set, and let $H = [h_1, h_2, \ldots, h_K] \in \mathbb{R}^{d \times K}$ be the codebook learned by LSR. Then the problem of LSR is formulated as follows:

$$\begin{aligned} &\min_{H,M} \|X - HM\|_F^2 + \lambda \sum_i \|m_i\|_1 + \frac{\beta}{2} \sum_{ij} \|m_i - m_j\|^2 S_{ij} \\ &= \min_{H,M} \|X - HM\|_F^2 + \lambda \sum_i \|m_i\|_1 + \beta \cdot trace(MLM^T) \\ &s.t. \ \|h_m\|^2 \leq 1 \end{aligned} \tag{2}$$

where the m_i denotes the Laplacian sparse representation of the i-th feature vector x_i and $M = [m_1, m_2, \ldots, m_N] \in \mathbb{R}^{K \times N}$. The λ is a regularization parameter with respect to the sparsity and β is the similarity constraint. The Laplacian

matrix is $L = D - S$ and the $D_{ii} = \sum_j S_{ij}$ denotes the degree of the i-th node. We use the histogram intersection $I(x_i, x_j)$ to compute the similarity matrix S.

$$I(x_i, x_j) = \sum_{l=1}^{d} \min(x_{i_l}, x_{j_l}) \tag{3}$$

where x_{i_l} is the l-th entry of the d dimensional feature vector x_i. Let $S_{ij} = S_{ji} = I(x_i, x_j)$ if x_i is the k nearest neighbor of x_j, where $i \neq j$ and $k = 5$. Otherwise let $S_{ij} = 0$.

The problem defined by Eq. (2) is not convex if we optimize H and M concurrently. However, when one of H and M is fixed, the optimization for the other one is convex. Thus, H and M are optimized alternately following the work [3].

Firstly, we fix the codebook H, then each element in the sparse representation set M can be optimized individually. If we want to compute each m_i, the other sparse representations m_j, where $j \neq i$, should be fixed. Then the feature sign search algorithm [7] is used to solve the following optimization to produce m_i:

$$\min_{m_i} \|x_i - Hm_i\|^2 + \beta(m_i^T(ML_i) + (ML_i)^T m_i - m_i^T L_{ii} m_i) + \lambda\|m_i\|_1 \tag{4}$$

where the i-th column of the Laplacian matrix L is L_i, and the (i, i)-th element of L is L_{ii}. M is the initialized sparse representation set and should be updated after the optimization of m_i, where m_i denotes the i-th column of M.

When we fix the sparse representation set M, the learning of codebook H can be defined as:

$$\min_{H} \|X - HM\|_F^2 \qquad s.t.\ \|h_m\|^2 \leq 1 \tag{5}$$

The Lagrange dual proposed in [7] is utilized to solve the Eq. (5).

In our approach, we randomly extract some features to learn the codebook H by optimizing Eqs. (4) and (5) iteratively, which is an offline learning task. After we obtain the codebook H, we only need to solve Eq. (4) to learn the sparse representation for each feature in the spatio-temporal features set X.

2.3 Maximum Pooling

Through the Laplacian sparse representation method, we can obtain a set of sparse representations of a query video, denoted by $\{m_1, m_2, \ldots, m_n\}$. Then we utilize the maximum pooling over this sparse representation set to obtain a K dimensional feature vector ρ which is used to characterize the video. Let ρ_j denote the j-th element of ρ, then the maximum pooling can be defined as:

$$\rho_j = \max\{|m_{1_j}|, |m_{2_j}|, \ldots, |m_{n_j}|\}, j \in \{1, \ldots, K\} \tag{6}$$

where m_{i_j} is the j-th element of the vector m_i. Each column of the codebook H represents a basic pattern of the feature space. With the maximum pooling method, we reserve the strongest response to each basic pattern and produce a feature vector with K dimensions for a query video, which also reduces the influence of irrespective information and improves the robustness of our approach.

2.4 Abnormal Activity Detection via SVM

The detection of abnormal activity in a video is a two-class classification problem. In our approach, we utilize the support vector machine (SVM) [2] with a radial basis function (RBF) kernel to solve this problem. The kernel trick used in SVM can map the linearly inseparable features into the high-dimensional space, which can make features linearly separable in the new space. After we obtain the feature vector ρ, we use the SVM to classify ρ as normal or abnormal.

3 Experiments

In this section, we conduct experiments on the UMN dataset [9] and the Hockey dataset [10] to test our approach.

3.1 Experimental Results on the UMN Dataset

This dataset contains 7740 frames with the crowded escape activities in 3 scenes. And the resolution is 320×240 pixels. We cut the whole video into 258 clips, including 196 normal video clips and 62 abnormal video clips. All video clips are divided into 5 subsets to test our approach for the global abnormal activity detection with 5-fold cross validation. In addition, the size of codebook in this experiment is set to 1000.

We compare our approach with three different methods on the UMN dataset, including social force model [9], optical flow [9], and streakline representation method [8]. The ROC curves are shown in Fig. 3 and quantitative comparisons are reported with the area under ROC curve (AUC) in Table 1.

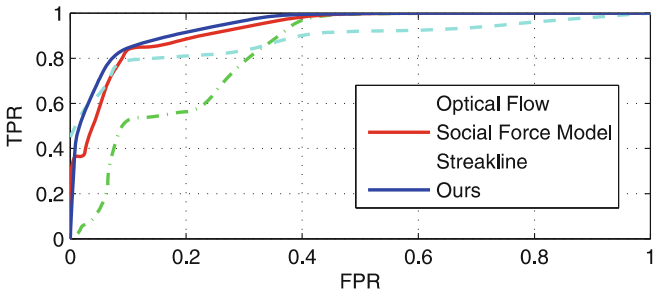

Fig. 3. The ROC curves on the UMN dataset

As reported in Table 1, the AUC of our approach is 0.971, which outperforms the optical flow method, streakline representation method and social force model. Experimental results on the UMN dataset have proved that our approach can perform effectively in the global abnormal activity detection.

Table 1. The performance comparisons on the UMN dataset with AUC

Method	Social Force Model [9]	Optical Flow [9]	Streakline [8]	Ours
AUC	0.96	0.84	0.90	0.971

We analyze the influence on our approach of parameters λ and β in Eq.(2). With the increase of λ, the sparse representation of feature is more sparse. We set λ among $0.3 \sim 0.4$ in the UMN dataset and find that $\lambda = 0.32$ performs best. The parameter β influences the similarity constraint in sparse representation generation. We set β among $0.1 \sim 0.3$ and find that our approach achieves good performance when $\beta = 0.1$.

3.2 Experimental Results on the Hockey Dataset

The Hockey dataset is composed of 1000 video clips with a resolution of 360×288 pixels, including 500 clips with fight events among athletes and 500 normal clips. Each video clip contains 50 frames. We test the proposed approach for the local abnormal activity detection on this dataset with 5-fold cross validation.

We present the comparisons among our approach, methods based on BoW model [10] and method using general sparse coding [12] in Table 2.

Table 2. The comparisons on the Hockey dataset with 5-fold cross validation

Visual Words	HOG with BoW [10]	HOF with BoW [10]	MoSIFT with BoW [10]	MoSIFT with sparse coding [12]		Ours	
	Accuracy	Accuracy	Accuracy	Accuracy	AUC	Accuracy	AUC
50	87.8 %	83.5 %	87.5 %	90.9 %	0.951	91.8 %	0.955
100	89.1 %	84.3 %	89.4 %	92.6 %	0.958	93.0 %	0.962
150	89.7 %	85.9 %	89.5 %	93.4 %	0.963	93.9 %	0.966
200	89.4 %	87.5 %	90.4 %	94.1 %	0.971	94.7 %	0.974
300	90.8 %	87.2 %	90.4 %	94.1 %	0.968	94.8 %	0.976
500	91.4 %	87.4 %	90.5 %	94.3 %	0.971	95.1 %	0.979
1000	91.7 %	88.6 %	90.9 %	94.0 %	0.967	95.3 %	0.978

According to Table 2, the approach using general sparse coding [12] gains a higher prediction accuracy than approaches based on BoW model, which indicates that the sparse coding method achieves better performance than BoW model in the feature quantization. Our approach obtains a higher prediction accuracy than the other approaches, which indicates that HNF descriptor is discriminative. Furthermore, the AUC of our approach is larger than approach

using general sparse coding, which indicates that Laplacian sparse representation method produces less feature quantization error because of the preserving of similarity among features. The experimental results indicate that our approach has achieved promising performance in the detection of local abnormal activity.

We also apply the same parameter analysis as the UMN dataset on this dataset. We find that our approach performs best when $\lambda = 0.3, \beta = 0.1$.

4 Conclusion

In this paper, we propose an approach based on spatio-temporal feature and Laplacian sparse representation to detect abnormal activity in a video. We use the HNF descriptor to characterize the appearance and action of a query video after the detection of space-time interest points. Then the Laplacian sparse representation and maximum pooling method are applied to obtain a more descriptive feature vector. With the introduction of similarity matrix in the LSR, we preserve the similarity among spatio-temporal features, which improves the accuracy and robustness in feature quantization. Experimental results on the UMN dataset and the Hockey dataset demonstrate that our approach can achieve satisfactory performance in the detection of both global abnormal activity and local abnormal activity.

Ackonwledgements. This research is partly supported by NSFC, China (No: 61273258) and 863 Plan, China (No. 2015AA042308).

References

1. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
2. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
3. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely-laplacian sparse coding for image classification. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3555–3561. IEEE (2010)
4. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
5. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2–3), 107–123 (2005)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
7. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in neural information processing systems, pp. 801–808 (2006)
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, 2009. CVPR 2009, pp. 935–942. IEEE (2009)
10. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011, Part II. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011)
11. de Souza, F.D.M., Chávez, G.C., do Valle, E., Araujo, D.A., et al.: Violence detection in video using spatio-temporal features. In: 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 224–230. IEEE (2010)
12. Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L.: Violent video detection based on mosift feature and sparse coding. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3538–3542. IEEE (2014)