ABNORMAL EVENT DETECTION USING SPATIO-TEMPORAL FEATURE AND NONNEGATIVE LOCALITY-CONSTRAINED LINEAR CODING

*Yu Zhao, Lei Zhou, Keren Fu, Jie Yang**

Institution of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China

ABSTRACT

In this paper, an approach using the spatio-temporal feature and nonnegative locality-constrained linear coding (NLLC) is proposed to detect abnormal events in videos. This approach utilizes position-based spatio-temporal descriptors as the low-level representations of a video clip. Each descriptor consists of the position information of a space-time interest point and an appearance feature vector. To obtain the high-level video representations, the nonnegative locality-constrained linear coding is adopted to encode each spatio-temporal descriptor. Then, the max pooling integrates all NLLC codes of a video clip to produce a feature vector. Finally, the support vector machine (SVM) is employed to classify the feature vector as abnormal or normal. Experimental results on two datasets have demonstrated the promising performance of the proposed approach in the detection of both global and local abnormal events.

Index Terms— abnormal event detection, spatio-temporal feature, nonnegative locality-constrained linear coding (NLLC), max pooling.

1. INTRODUCTION

With the increasing applications of computer vision techniques on intelligent surveillance systems, the abnormal event detection in videos has become an attractive research task. The abnormal events can be defined as unordinary activities such as the crowd disturbances, people fights. Depending on the scale of unordinary areas in scenes [1], the abnormal events can be divided into two classes: the global abnormal event (GAE) and the local abnormal event (LAE). For the GAE, the abnormal event happens in the entire scene, whereas the potential areas of the LAE are some local regions of the scene.

To detect the GAE in a video, Wu et al. [2] introduce the chaotic invariant of Lagrangian particle trajectories to characterize the events. In [3], the social force model based on the particle advection is used to analyze the human behaviour, and the abnormal event is detected by a Latent Dirichlet Allocation (LDA) model. On the LAE detection, the Mixture of Dynamic Texture (MDT) [4] is utilized to model the normal activities and outliers under this model are classified as abnormal. In [5], the Mixture of Probabilistic Principal Component Analysis (MPPCA) is employed to learn the motion pattern and the Markov Random Field (MRF) model is used to detect the abnormal event.

The abnormal event detection is influenced by the resolution of video, occlusion and camera movement. In order to overcome these disadvantages, Cong et al. [1] use the Multi-scale Histogram of Optical Flow (MHOF) descriptors as the low-level representations of a video clip. Then the sparse coding (SC) [6] is applied to produce high-level video representations, which has achieved promising performance in both GAE and LAE problems. Since neither human body tracking nor foreground segmentation is required, the approach based on sparse coding is robust to the scale of motion, cluttered background and viewpoint.

However, the discrimination of low-level descriptors has impact on the performance of sparse coding method. The MHOF descriptor adopted by [1] just provides the appearance information, ignoring the position information of descriptors. In [7], Liu et al. indicate that the location information of descriptors plays an important role in the image representation. Therefore, this paper proposes to combine the coordinate information of an interest point with the appearance information to form a position-based spatio-temporal descriptor for the abnormal event detection. On the other hand, the input descriptor may be assigned to a few visual words which are not near to the descriptor because of the over complete codebook of sparse coding[1]. This means that the results of sparse coding may be sparse but not local. In [8], Yu et al. point out that the locality is more essential than the sparsity and the locality can result in the sparsity of the coding coefficient. Moreover, the nonnegative coefficients are more accordant to the nonnegativity of neural firing rates [9]. Thus, our approach utilizes the nonnegative locality-constrained linear coding (NLLC) [10] to encode the spatio-temporal descriptors. The NLLC scheme introduces the locality and nonnegativity constraints to make the coding coefficients descriptive and robust.

The contributions of this paper are: (1) we introduce a new spatio-temporal descriptor containing both the position and appearance information. (2) we utilize the NLLC to detect abnormal events in videos and introduce an approximated implementation of this coding scheme.

*Corresponding author: Jie Yang, jieyang@sjtu.edu.cn

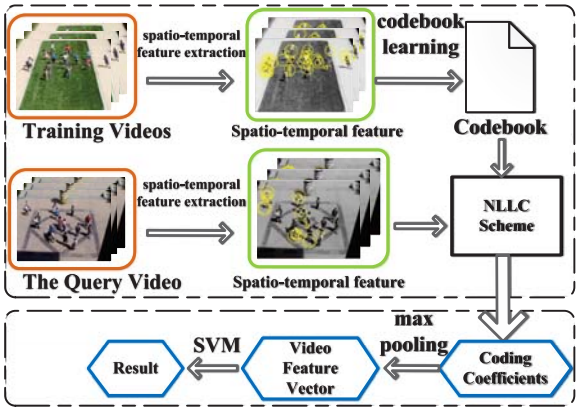

Fig. 1. The overview of the proposed approach.

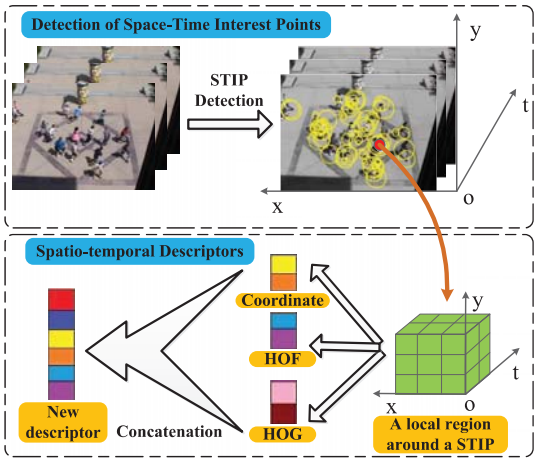

Fig. 2. The extraction of the spatio-temporal feature.

2. METHODOLOGY

This paper proposes an approach based on the position-based spatio-temporal feature and NLLC scheme to detect both global and local abnormal events in videos. The overview of the proposed approach is illustrated by the Fig.1. Firstly, the spatio-temporal descriptors are extracted as the low-level video representations. Then we apply the NLLC scheme to encode all descriptors of a video to obtain the high-level video representations. After that, the max pooling method is operated over the entire coding coefficient set of a video to produce a discriminative feature vector. Finally, the feature vector is classified as abnormal or normal by the SVM [11].

2.1. Spatio-temporal feature

To produce discriminative low-level representations of a video clip, we introduce the position-based spatio-temporal descriptors. The extraction of the descriptors consists of the following two procedures, illustrated by Fig.2.

(1) Detection of space-time interest points. The interest points are located at local regions with intense variation of image intensity in the spatio-temporal domain. This paper adopts the 3D Harris detector [12] to detect the space-time interest point (STIP). Firstly, let $f_{st} : \mathcal{R}^2 \times \mathcal{R} \mapsto \mathcal{R}$ model a video clip. Then the f_{st} is converted into the linear space and we obtain L_{st}. By applying the convolution of a spatio-temporal Gaussian kernel with the second-moment matrix of L_{st}, we can produce the detection matrix Ψ_{st}. The significant eigenvalues of Ψ_{st} indicate the local patches which have intense variation of the f_{st}. Thus, the following response function can be utilized to find the space-time interest points.

$$\begin{aligned} H_{st} &= \lambda_1\lambda_2\lambda_3 - \rho \cdot (\lambda_1 + \lambda_2 + \lambda_3)^3 \\ &= \lambda_1^3(\mu\nu - \rho \cdot (1 + \mu + \nu)^3) \end{aligned} \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the eigenvalues of the matrix Ψ_{st}. The function $H_{st} \geq 0$, with $\rho(1 + \mu + \nu)^3 \leq \mu\nu$, where $\mu = \lambda_2/\lambda_1, \nu = \lambda_3/\lambda_1$. Then space-time interest points can be detected by searching the local positive maximum values of H_{st}. And the spatio-temporal coordinate of each interest point is denoted as $w_{st} = (x, y, t)^T$.

(2) Extraction of the spatio-temporal descriptor. To produce the position-based spatio-temporal descriptor, we concatenate the coordinate of an interest point with the Histograms of Oriented Gradients and Optical Flow (HNF) [13]. As illustrated in Fig.2, a local region around an interest point is divided into a grid with $3 \times 3 \times 2$ spatio-temporal patches. In each patch, the 4-bin Histograms of Oriented Gradients (HOG) descriptor and the 5-bin Histograms of Optical Flow (HOF) descriptor are computed respectively. Then we combine the HOG with the HOF to obtain the HNF descriptor which has 162 dimensions. Motivated by [7], the position-based spatio-temporal descriptor with 165 dimensions is defined as:

$$\phi = [w_{st}^T, HNF^T]^T = [x, y, t, HNF^T]^T \tag{2}$$

The descriptor ϕ is normalized by the ℓ_2 normalization.

2.2. The NLLC scheme for spatio-temporal descriptors

In this paper, the proposed approach utilizes the NLLC scheme [10] to produce discriminative high-level video representations for the abnormal event detection. Let $\Phi = [\phi_1, \phi_2, \ldots, \phi_S](\phi_i \in \mathcal{R}^{d\times 1})$ denote the descriptors set of a video clip, then the NLLC problem is formulated as follows:

$$\begin{aligned} &\min_M \sum_{i=1}^{S} \|\phi_i - Cm_i\|^2 + \lambda\|d_i \odot m_i\|^2 \\ &s.t.\ \mathbf{1}^T m_i = 1, m_i \geq 0, \forall i \end{aligned} \tag{3}$$

where the $C = [c_1, c_2, \ldots, c_\ell] \in \mathcal{R}^{d\times\ell}$ is a pre-learnt codebook. And the codebook learning of NLLC will be discussed

in the subsection 2.3. The $M = [m_1, m_2, \ldots, m_S] \in \mathcal{R}^{\ell \times S}$ denotes the NLLC coefficient set of a video clip. The constraint $\mathbf{1}^T m_i = 1$ makes the NLLC coefficient meet the shift-invariant property. λ is the parameter which makes the trade-off between the reconstruction error and the locality. The operator $\odot$ is the element-wise multiplication. d_i denotes the exponential Euclidean distance between the descriptor and each visual word. It is defined as:

$$d_i = [\exp(\|\phi_i - c_1\|^2/\sigma), \ldots, \exp(\|\phi_i - c_\ell\|^2/\sigma)]^T \quad (4)$$

where σ is the scale parameter of the Gaussian function.

In order to speed up the feature quantization procedure, the proposed approach adopts an approximated implementation of NLLC for fast coding. Note that the locality-constraint in Eq.(3) selects several local visual words which are near to each descriptor to quantize it [14]. Moreover, the nonnegativity constraint transforms the coding coefficient into a vector with a few positive elements and many zero elements. Thus, we can first solve the following optimization without considering the locality and nonnegativity constraint.

$$\begin{aligned} &\min_{\widetilde{M}} \sum_{i=1}^{S} \|\phi_i - C_i \widetilde{m}_i\|^2 \\ &s.t.\ \mathbf{1}^T \widetilde{m}_i = 1, \forall i \end{aligned} \quad (5)$$

where the C_i is a subset of the codebook. We select the K nearest neighbors of x_i from the codebook to produce the bases set C_i. In this paper, the K is set to 5. The fast encoding algorithm proposed by [14] is utilized to solve the Eq.(5) to produce an approximated code $\widetilde{m}_i$ for each descriptor.

After producing the approximated code $\widetilde{m}_i$, the nonnegativity property should be imposed. Inspired by [15], the following process is employed to make the $\widetilde{m}_i$ meet the nonnegativity constraint.

$$\widetilde{m}_{ij} \leftarrow \frac{\widetilde{m}_{ij} + |\min\{\widetilde{m}_{ij^\dagger}, 0\}|}{\sqrt{\sum_j (\widetilde{m}_{ij} + |\min\{\widetilde{m}_{ij^\dagger}, 0\}|)^2}} \quad (6)$$

where $j^\dagger = \underset{j}{\operatorname{argmin}}\, \widetilde{m}_{ij}$, and the $\widetilde{m}_{ij}$ is the j-th entry of the coefficient vector $\widetilde{m}_i$.

Through the two processes mentioned above, we accelerate the coding procedure and obtain an approximated implementation of the NLLC scheme.

2.3. Codebook learning for the NLLC scheme

The codebook is assumed to be given in all above discussion. In this subsection, we discuss the codebook learning of NLLC. Each visual word in the codebook represents a basic pattern of the spatio-temporal feature space. In this paper, we first extract descriptors from the training video clips as a subset Θ. Then the $\Theta = [\theta_1, \theta_2, \ldots, \theta_N] (\theta_i \in \mathcal{R}^{d \times 1})$ is used to learn the codebook by solving the following optimization.

$$\begin{aligned} &\min_{C,H} \sum_{i=1}^{N} \|\theta_i - C h_i\|^2 + \lambda \|d_i \odot h_i\|^2 \\ &s.t.\ \mathbf{1}^T h_i = 1, h_i \geq 0, \forall i \\ &\quad \|c_j\|^2 \leq 1, \forall j \end{aligned} \quad (7)$$

The Eq.(7) is a joint optimization problem with respect to the C and H, where $H = [h_1, h_2, \ldots, h_N] \in \mathcal{R}^{\ell \times N}$ is the code set. We first utilize the K-Means clustering to learn an initial codebook C_{init} from the subset Θ. Then the incremental codebook learning algorithm [14] is utilized to solve the Eq.(7) to produce the codebook C.

2.4. Video representation based on max pooling

Through the NLLC scheme, the coding coefficient set M of a video clip is produced. Then all NLLC codes of a video should be integrated together to produce a global video feature vector α. In [16], Gao et al. indicate that max pooling can restrain noises of coding coefficients, which outperforms the average pooling. Thus, the proposed approach adopts the max pooling to integrate NLLC codes of a video. Let α_j denote the j-th entry of α, and the max pooling can be formulated as:

$$\alpha_j = \max\{|m_{1j}|, |m_{2j}|, \ldots, |m_{Sj}|\} \quad (8)$$

where m_{ij} is the j-th element of the code m_i. Then the video feature α is normalized through the ℓ_2 normalization.

In summary, the proposed approach extracts the position-based spatio-temporal descriptors to describe a video clip. Then the NLLC scheme generates high-level representations corresponding to these descriptors. By applying max pooling method, the discriminative and robust video feature is produced. Then the video feature is classified as normal or abnormal via the SVM.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed approach on the GAE and the LAE problems, we conduct experiments on the UMN [17] and the USCD ped 1 [4] datasets. The size of codebook is set to 500 in all experiments.

3.1. Global abnormal event detection

The proposed approach is tested on the UMN dataset for the detection of global abnormal event. The UMN dataset contains crowded escape events in three different scenes, with a 320×240 resolution. The whole dataset is divided into 258 video clips, including 50 abnormal clips and 208 normal video clips. We choose 10 abnormal video clips and 10 normal video clips as the training set to train the SVM and the rest video clips make up the test set.

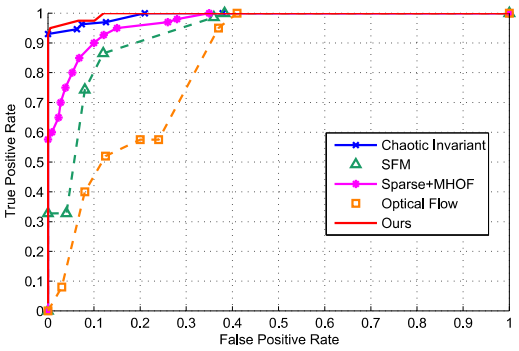

Fig. 3. The ROC curves on the UMN dataset.

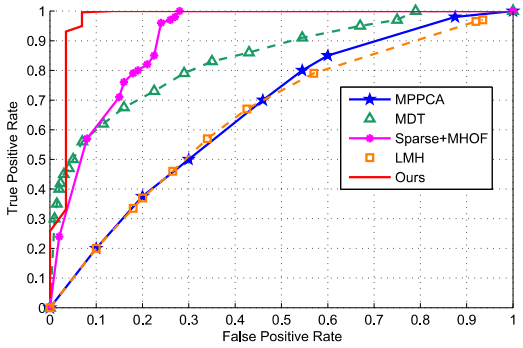

Fig. 4. The ROC curves on the USCD ped 1 dataset.

Table 1. The quantitative comparisons on the UMN dataset.

Method	Chaotic Invariant	SFM	Sparse+ MHOF	Optical Flow	Ours
AUC	0.99	0.96	0.978	0.84	0.996

Table 2. The EER on the USCD ped 1 dataset.

Method	MPPCA	MDT	Sparse+ MHOF	LMH	Ours
EER	40%	25%	19%	38%	6.9%

we compare the proposed approach with different methods, including Chaotic Invariant [2], Social Force Model (SFM) [3], Sparse Reconstruction [1], and Optical Flow [3]. The ROC curves on this dataset are shown in Fig.3 and the area under the ROC cure (AUC) is reported in the Table 1.

As reported in the Table 1, the AUC of our approach is 0.996, which outperforms the other methods. Compared with MHOF descriptor [1], the position-based descriptor adopted by our method has achieved better performance. It demonstrates that the position information has improved the discrimination of the new descriptor. Furthermore, the promising performance shows that the NLLC scheme can produce more descriptive hight-level video representations because of introducing the locality constraint into the coding phase. The results on the UMN dataset indicate that the proposed approach can perform effectively in the detection of global abnormal events.

The parameters σ in Eq.(4) and λ in the Eq.(7) influence the learning of the coodbook. The proposed approach is tested on the UMN dataset while σ and λ are respectively set to $1 \sim 50$. Finally, the approach achieves the best performance with the $\sigma = 10$ and $\lambda = 10$.

3.2. Local abnormal event detection

We test the proposed approach on the USCD ped 1 dataset for the local abnormal event detection. This dataset consists of 36 abnormal video clips and 34 normal video clips, with a 158×238 resolution. 5 abnormal clips and 5 normal clips are selected as the training set to train the SVM and the reset video clips are used to test the proposed approach. The results are illustrated in the Fig.4 with the ROC curves and the quantitative comparisons are reported in the Table 2 with the equal error rate (EER). The proposed approach is compared with different methods, including MPPCA [5], MDT [4], Sparse Reconstruction [1], and LMH [18].

According to the Table 2, the EER of the proposed approach is 6.9%, which is lower than the other methods. It proves that the proposed position-based descriptor is descriptive enough to model the position and appearance information. Compared with the method based on sparse coding [1], the proposed approach achieves a better ROC curve and a lower EER, indicating that the NLLC scheme produces less quantization error and generates more robust coding coefficients than the sparse coding. The results on this dataset demonstrate that the proposed approach can also achieve promising performance in the local abnormal event detection.

We also apply the parameter analysis on this dataset and find that our approach performs best when $\sigma = 10, \lambda = 0.1$.

4. CONCLUSION

In this paper, we propose an approach using the position-based spatio-temporal feature and the NLLC scheme to detect abnormal events in videos. The position-based descriptors provide the location and appearance information of local motion patterns. With the introduction of locality and non-negativity, the NLLC scheme can produce more discriminative high-level video representations. Then the max pooling method integrates these NLLC codes of a video clip to generate a video feature vector for the final classification. The experimental results have shown that the proposed approach can perform effectively in the detection of both GAE and LAE.

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

5. REFERENCES

[1] Yang Cong, Junsong Yuan, and Ji Liu, "Sparse reconstruction cost for abnormal event detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3449–3456.

[2] Shandong Wu, Brian E Moore, and Mubarak Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2054–2060.

[3] Ramin Mehran, Akira Oyama, and Mubarak Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 935–942.

[4] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1975–1981.

[5] Jaechul Kim and Kristen Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2921–2928.

[6] Jianchao Yang, Kai Yu, Yihong Gong, and Tingwen Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[7] Ming Liu, Shuicheng Yan, Yun Fu, and Thomas S Huang, "Flexible xy patches for face recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2113–2116.

[8] Kai Yu, Tong Zhang, and Yihong Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223–2231.

[9] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] Yuanbo Chen and Xin Guo, "Learning non-negative locality-constrained linear coding for human action recognition," in *Visual Communications and Image Processing (VCIP), 2013*. IEEE, 2013, pp. 1–6.

[11] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[12] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[13] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[14] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[15] Peina Liu, Guojun Liu, Maozu Guo, Yang Liu, and Pan Li, "Image classification based on non-negative locality-constrained linear coding," *Acta Automatica Sinica*, vol. 41, no. 7, pp. 1235–1243, 2015.

[16] Shenghua Gao, Ivor Wai-Hung Tsang, Liang-Tien Chia, and Peilin Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3555–3561.

[17] "Unusual crowd activity dataset of minnesota university from," `http://mha.cs.umn.edu/movies/crowd-activity-all.avi`.

[18] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and David Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 3, pp. 555–560, 2008.