Anomaly Detection in Crowded Scenes by SL-HOF Descriptor and Foreground Classification

Siqi Wang, En Zhu, Jianping Yin
College of Computer
National University of Defense Technology
Changsha, China
Email: wangsiqi10c@gmail.com, {enzhu, jpyin}@nudt.edu.cn

Fatih Porikli
College of Engineering and Computer Science
Australian National University
Canberra, Australia
Email: fatih.porikli@anu.edu.au

***Abstract*—With the widespread use of surveillance cameras, massive video data analysis has become an extremely labor-intensive work. In this paper, we propose an efficient approach to detect video anomaly in crowded scenes based on Spatially Localized Histogram of Optical Flow (SL-HOF) descriptor and foreground classification. For motion description, the new SL-HOF descriptor can not only preserve classic HOF descriptor's favorable capability of characterizing the motion velocity and direction of foreground in crowded scene, but also depicts the spatial distribution of optical flow, which implicitly encodes the structure and local motion information of foreground objects in videos. SL-HOF is shown to significantly outperform other classic video descriptors. To further boost the performance of anomaly localization, we then introduce Robust PCA based foreground classification to discriminate anomalous foreground texture. Instead of computationally expensive approaches like l_1-norm Sparse Coding, we adopt classic one-class SVM (OCSVM) to model normal video events and detect outliers (anomaly). Our experiments on the challenging UCSD datasets show our approach can achieve state-of-the-art results when compared to existing video anomaly detection methods.**

I. Introduction

Video anomaly detection, which plays a center role in smart video surveillance technology, has drawn increasing interest from academia and industry due to its substantial potential in liberating human beings from long-time tedious work in monitoring possibly as many as hundreds of surveillance screens. Various applications of video anomaly detection can be found in realms such as public security and city management.

Possibilities always come with challenges. The most fundamental challenge in video anomaly detection lies in the definition of "normal/abnormal event" is not as straightforward as the "face" or "pedestrians". Early work like [1] attempted to explicitly describe certain "abnormal" events or behaviors. However, such methods are severely limited to certain occasions, not to mention that prior knowledge of anomaly is often unavailable. Therefore, recent works tend to consider video anomaly detection as an "outlier detection" problem [2]. In such methods, only normal video events are modeled and "anomaly" is considered to be those events diverting significantly from normal events, the idea of which is also followed in this paper. Another key challenge comes from video representation. Object tracking and trajectory analysis [3] [4] [5] [6] seem to be a natural idea to capture high level feature and it works well in uncrowded scenes. Unfortunately, tracking based methods perform poorly in crowded scenes due to frequent occulusions and complex foreground motion. Thus, robust low level features are proposed to deal with crowded scene anomaly detection. Mahadevan *et al.* [7] proposed the challenging UCSD datasets with crowded scenes and used Mixture of Dynamic Texture (MDT) for a joint modeling of foreground appearance and dynamics. Cong *et al.* [8] proposed a popular Multi-scale Histogram of Optical Flow (MHOF) descriptor. Roshtkhari *et al.* [9] represent spatio-temporal video volume by classic HOG descriptor, while Zhao *et al.* [10] combined 3D HOG and HOF for video representation in anomaly detection. Proposed by Kratz *et al.* [11], 3D gradient is adopted by [12] [13]. The third challenge is video event modeling. Sparse Coding is a representative category among those methods. Cong *et al.* [8] reconstructed a new event by a dictionary, which consists of representative normal events, by solving a l_1-norm optimization problem. Lu *et al.* [12] learned multiple small fixed-size sparse combinations to enable a high-speed detection process. In addition to Sparse Coding, Antic *et al.* [14] extracted a set of foreground hypotheses to jointly explain all foreground pixels in testing videos. Chen *et al.* [15] proposed to use Gaussian Process Regression (GPR) and hierarchical feature representation to detect video anomaly. Zhang *et al.* [13] modeled appearance by Support Vector Data Description (SVDD), while Saligrama *et al.* [16] and Mahadevan *et al.* [7] both adopted probabilistic models to describe normal video events.

In this paper, we address video anomaly detection from crowded scenes based on the proposed SL-HOF and foreground classification. The rest of paper is organized as follows: In Sec. II, we discuss how to represent video event with SL-HOF descriptor in terms of motion and analyze the underlying reasons why SL-HOF based representation is effective. In Sec. III we introduce foreground classification as a supplement to SL-HOF based representation to model video foreground appearance on pixel intensity level,. Robust PCA is used to extract video foreground and generate textural features of foreground objects in normal video events. Finally, both SL-HOF features and textural features of normal video events are modeled by OCSVM. Experiments and results on UCSD datasets are shown in Sec. IV. Sec. V concludes the paper.

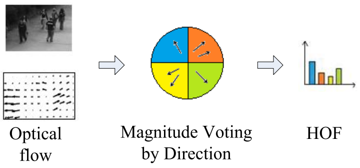

Fig. 1: The calculation procedure of HOF descriptor.

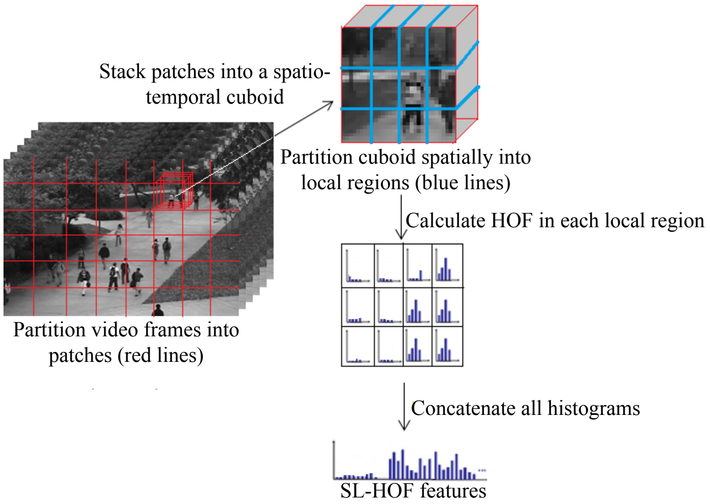
Stack patches into a spatio-temporal cuboid
Partition cuboid spatially into local regions (blue lines)
Calculate HOF in each local region
Partition video frames into patches (red lines)
Concatenate all histograms
SL-HOF features

Fig. 2: The calculation procedure of SL-HOF descriptor.

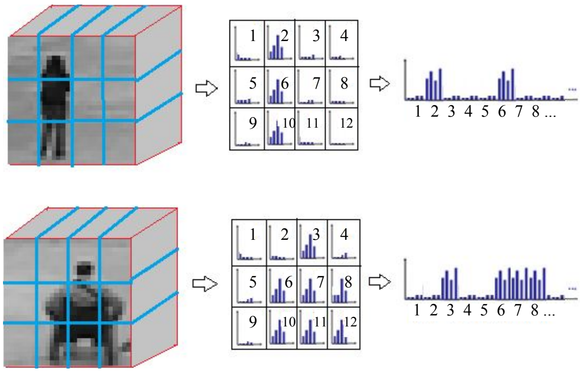

Fig. 3: Foreground structural information embedded in SL-HOF based video representation.

II. SL-HOF Descriptor

In this section, optical flow and classic Histogram of Optical Flow (HOF) descriptor is briefly reviewed. Then we present the SL-HOF video descriptor and explain the reasons of SL-HOF's effectiveness.

A. *Optical flow and HOF*

Optical flow is calculated by estimating the motion velocity and direction of each pixel from two consecutive video frames. Optical flow has been widely used in video anomaly detection due to its powerful capability of describing motions in video. To represent motion, HOF descriptor is a frequently used descriptor. To calculate HOF, the optical flow magnitude of each pixel in region of interest (e.g. spatio-temporal cuboid from video) is voted into D bins by their optical flow directions to obtain a D-bin histogram (See Fig. 1) as a HOF feature.

B. *SL-HOF*

Classic HOF is not discriminative enough to detect different complex anomaly, therefore SL-HOF is proposed. The calculation process of a SL-HOF video representation is shown as Fig. 2: Firstly, each video frame is partitioned into non-overlapping $M \times N$ patches with equal size, and patches at the same spatial location in consecutive d video frames are stacked as a spatio-temporal cuboid, which is a standard practice in video anomaly detection. Each spatio-temporal cuboid is assumed to be a video event with one or several foreground objects. d is usually small in SL-HOF (e.g. $d = 5$ in our configuration). This aims to characterize the foreground motion in a small temporal interval, in which the foreground does not suffer from drastic change in spatial location. Then, a spatio-temporal cuboid is partitioned spatially into $m \times n$ local 3D regions. Unlike cell-based 3D HOG [17], we do not partition the spatio-temporal cuboid temporally because it will dampen the motion statistics in each region and increase feature dimension. Next, histograms of optical flow are calculated from each region rather than the entire cuboid, which differentiates SL-HOF from HOF. Finally, all histograms are concatenated to obtain a SL-HOF representation $\mathbf{v} = [\mathbf{h}_1^{\mathbf{T}}, \mathbf{h}_2^{\mathbf{T}}, ...\mathbf{h}_{m\times n}^{\mathbf{T}}]^{\mathbf{T}}$ as the motion summarization of the cuboid (or video event).

The operations done by SL-HOF descriptor are simple. However, our experiments show SL-HOF descriptor works surprisingly well in video anomaly detection (See Sec. IV-B). The following two properties of SL-HOF contribute to its sound performance: First of all, SL-HOF can depict the distribution of optical flow in video foreground objects, which implicitly encode the structural information of foreground objects into its vector representation. Consider an example of a walking man (normal event) and a man in the wheelchair (abnormal event) in Fig. 3. When the wheelchair moves at a close speed and direction to the walking man, HOF descriptor can be easily fooled since classic HOF extracted from the entire spatio-temporal cuboid wipes out the spatial location information of foreground when calculating histogram. By contrast, SL-HOF yields completely different feature vectors because the spatial distributions of optical flow for man and wheelchair are totally different, which is caused by their different structure. To be more specific, strong optical flow histograms (by ”strong” we mean at least one bin of the histogram has a large vote value) are found in region 2, 6 and 10 for the walking man, while strong optical flow histograms are observed in region 3, 6, 7, 8, 10, 11, 12 for man in the wheelchair (See Fig. 3). Therefore, localized HOF extraction can preserve the rough location information of optical flow during calculating histograms. Consequently, SL-HOF can implicitly embed the structural information of foreground objects through optical flow's spatial distribution.

Secondly, SL-HOF can capture local motion information

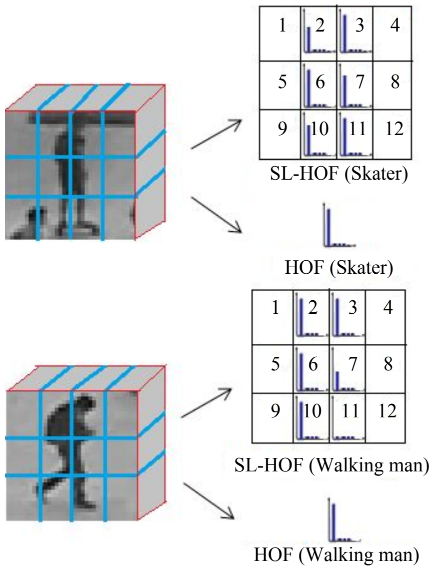

Fig. 4: Local motion information embedded in SL-HOF based video representation.

of foreground objects. Local motion information can be important for discriminating different foreground objects, which is illustrated by Fig. 4: A man (normal event) and a skater (abnormal event) both heading towards right. Two objects share very similar structure and appearance, and it will be difficult to discriminate them when the skater share a close speed with the walking man. When representing both objects with HOF descriptor, the difference will be minor since almost all pixels in both objects are moving towards right and their optical flow are accumulated into the same bin (The first bin in this example). However, it should be noted that the local motion of each body part is not consistent when human is walking. For example, human's two legs advance alternatively when walking. As shown in Fig. 4, the man's supporting leg in region 11 almost remain static while the other leg in region 10 is rapidly moving, thus leading to different histograms in region 10 and 11: region 10 can observe a strong optical flow histogram while region 11 cannot. However, since the skater moves as a whole on skateboard, both region 10 and 11 can observe a strong optical flow histogram. SL-HOF can capture such difference in local motions. Besides, in crowded scenes with severe occlusions, localized histogram extraction enables SL-HOF to describe the local motions of individual object parts while HOF obviously cannot. Thus, SL-HOF is more discriminative than HOF or MHOF. SL-HOFs extracted from the same spatial location are described by OCSVM.

III. Foreground Classification

In addition to motion anomaly, there exists appearance or texture anomaly that can be directly classified by eyes from the foreground of a single video frame, e.g., a bicycle on the pavement. Since SL-HOF has been used to characterize the motion of video foreground, we propose foreground classification as a supplement to classify anomalous foreground texture. The procedure is given in Fig. 5: The first step is to perform Robust PCA [18] on video sequences to obtain a sparse matrix E_t for video frame f_t. Afterwards, $p_{i,j}$, the probability that pixel $f_t^{i,j}$ belongs to foreground, is estimated by $p_{i,j} = 2/exp(-\lambda \cdot (E_t^{i,j})^2) - 1$, where $\lambda = 1/N_p$, N_p is the total number of pixels on a frame. Next, the probability map is binarized by threshold 0.5 to obtain the foreground. To generate normal texture samples, a fixed-size bounding box is generated by taking each foreground pixel as the box center. To filter out redundant boxes, only those boxes with $> 30\%$ pixels to be foreground are preserved. Non-maximum suppression and sampling are adopted to further reduce the number of boxes. Remaining boxes are selected as texture samples and described by HOG descriptor. Texture samples with centers at the same spatial location (patch) are collected for training OCSVM, which is applied to discriminate anomaly from testing texture samples. To control false alarms, the decision threshold of OCSVM is lowered.

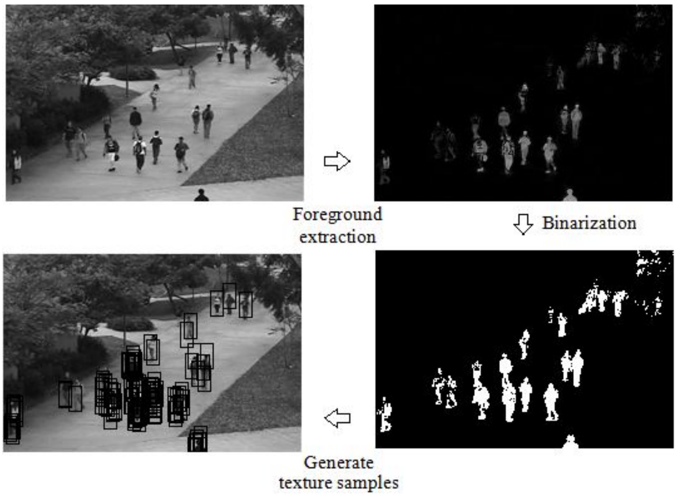

Fig. 5: Foreground extraction and texture samples generation.

IV. Experiments

In this section, we test the proposed approach on the most commonly-used UCSD datasets with crowded scenes. In Sec. IV-A, we introduce the adopted UCSD ped1 and ped2 datasets and the configuration of experiments. Sec. IV-B compares the proposed SL-HOF descriptor with other classic video descriptors. Sec. IV-C demonstrate the effect of combining SL-HOF representation and foreground classification. Detection results and comparison with other state-of-the-art methods on are given in Sec. IV-D. Equal Error Rate (EER), ROC Curve and Area under the Curve (AUC) under two commonly-used evaluation criteria, frame-level and pixel-level criteria from [7], are adopted for method comparison. All experiments are carried out under MATLAB 2015b on a PC with 32 GB RAM and 3.90 Ghz Intel i7 4790 processor.

A. Datasets and Experimental Configuration

UCSD ped1 dataset consists of 34 training video sequences and 36 testing video sequences with 200 158×238 pixel

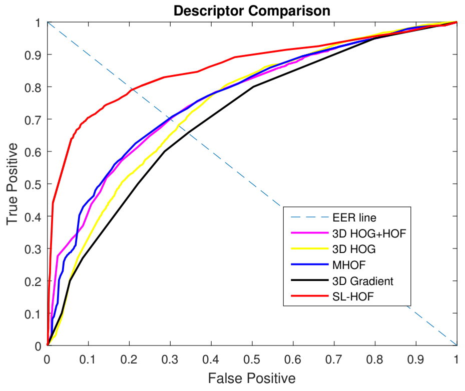

Fig. 6: Comparison of video descriptors.

frames per volume. UCSD ped2 dataset contains 16 training sequences and 12 testing video sequences with 240×360 video frames, and the numbers of which range from 120 to 180. All of the training volumes merely include normal events such as pedestrians walking on the pavement, while testing volumes contain abnormal events such as a skater and a vehicle on the pavement in crowded or uncrowded scenes.

The detection configuration is as follows: A local patch on a video frame is set to be 10×10, with consecutive $D = 5$ patches are stacked into a spatio-temporal cuboid. Features extracted from cuboids are assembled into the spatio-temporal basis from [8] to incorporate neighboring spatial and temporal correlation. For SL-HOF, the cuboid is partitioned into 7×8 regions to yield best performance, and PCA is adopted to reduce SL-HOF feature dimension. The video frames are resized into three scales for detection: 120×180, 100×150 and 80×60 for UCSD ped1 dataset, and 180×270, 120×180, 100×150 for UCSD ped2 dataset. Foreground classification is conducted on the original scale with patch size 21×21 and bounding box size 20×10. Both SL-HOF features of spatio-temporal cuboids and HOG features of texture samples are described by OCSVM [19] with Gaussian kernel. The parameters ν and σ are selected from $2^{-12}, 2^{-11}, ..., 2^{0}$ and $2^{-12}, 2^{-11}, ..., 2^{12}$ by 10-fold cross-validation, respectively.

B. Descriptor Comparison

In this section, the proposed SL-HOF descriptor is compared with the following frequently-used video descriptors in anomaly detection: MHOF [8], 3D HOG [20], 3D HOG+HOF [10], 3D Gradient [12]. For convenience, a single scale (100×150) detection is performed on UCSD ped1 dataset with frame-level evaluation criteria using different descriptors. Other configurations are the same as that in Sec. IV-A. The ROC curves obtained by different descriptors are given in Fig. 6 and the EERs are summarized in Tab. I: SL-HOF can improve detection performance significantly by approximately 10% EER reduction, and AUC is improved by 9% to 15%.

C. Combination of SL-HOF and Foreground Classification

In this section, we show the effect of combining SL-HOF with foreground classification (FC). Detection with SL-HOF only and detection with SL-HOF and FC are performed on UCSD ped1 dataset under more precise pixel-level evaluation criteria. As shown in Fig. 7 and Tab. II, foreground classification can enhance the anomaly localization performance by detecting texture anomaly, despite generating slightly more false alarms.

TABLE I: Comparison of video descriptors.

Descriptor	EER	AUC
MHOF	29%	76.45%
3D HOG	31%	73.98%
3D HOG+HOF	29%	76.28%
3D Gradient	33%	70.87%
SL-HOF	**21%**	**85.73%**

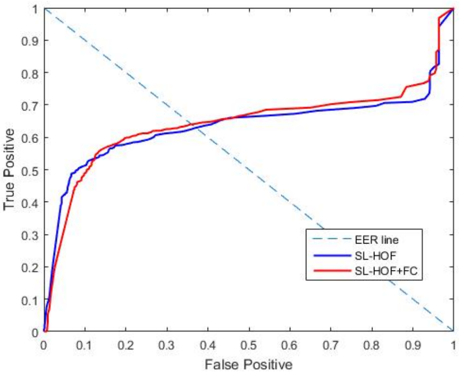

Fig. 7: Combining SL-HOF with FC.

TABLE II: Combining SL-HOF with FC.

	EER	AUC
SL-HOF	37 %	63.57 %
SL-HOF+FC	**35** %	**64.35**%

D. Method Comparison

In this section, we compare the proposed approach with state-of-the-art approaches in literature. For UCSD ped1 dataset, the following approaches are compared: Gaussian Process Regression (GPR) based method [15], Sparse Combination Learning (SC) [12], Mixture of Dynamic Texture (MDT) [7], Sparse Reconstruction Cost (SRC) [8], Social Force (SF) [21], Social Force and MPPCA (SF+MPPCA), Dense STC [9] and Adam *et al.* [22]. EERs, ROC Curves and AUCs of each methods are compared under both frame-level and pixel-level criteria (See Fig. 8, Fig. 9, Tab. III). From Tab. III, our approach yields comparable results with state-of-the-art methods under frame-level criteria while it evidently outperforms other methods under more precise pixel-level criteria.

As to results on UCSD ped2 dataset, the following state-of-the-art approaches are listed for comparison: Spatio-temporal Composition (STC) [23], motion and appearance cue (Zhang *et al.* [13]), Mixture of Dynamic Texture (MDT) [7], MPPCA

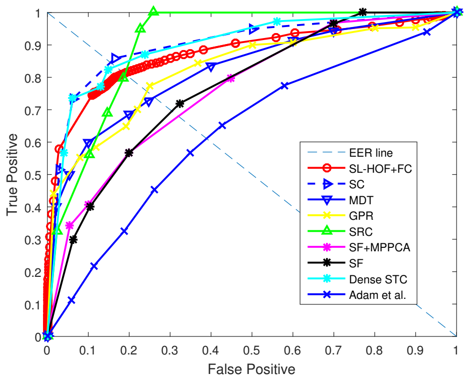

Fig. 8: Frame-level evaluation on UCSD ped1.

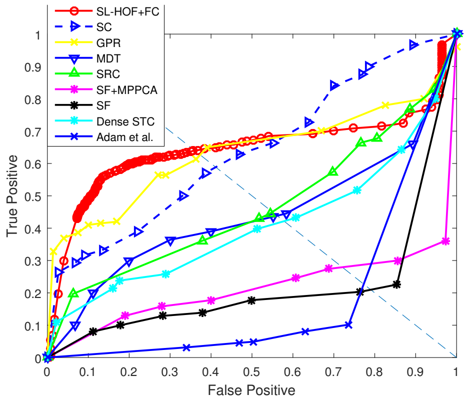

Fig. 9: Pixel-level evaluation on UCSD ped1.

TABLE III: Detection Results on UCSD ped1 dataset.

Method	EER(frame)	AUC	EER(pixel)	AUC
SL-HOF+FC	18%	87.45%	**35%**	**64.35%**
SC	**15%**	**91%**	40.3%	63.8%
GPR	23.8%	83.8%	37.3%	63.3%
SF	31%	67.5%	79%	19.7%
SF+MPPCA	32%	67%	71%	21.3%
MDT	25%	81.8%	55%	44.1%
SRC	19%	86%	54%	46.1%
Dense STC	16%	89%	58%	41.7%
Adam *et al.*	38%	65%	76%	13.3%

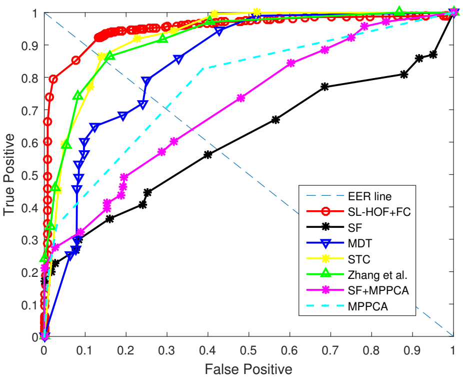

Fig. 10: Frame-level evaluation on UCSD ped2.

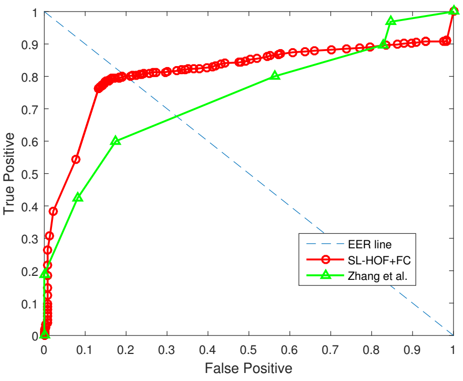

Fig. 11: Pixel-level evaluation on UCSD ped2.

TABLE IV: Detection Results on UCSD ped2 dataset.

Method	EER(frame)	AUC	EER(pixel)	AUC
SL-HOF+FC	**9%**	**95.07%**	**19%**	**81.04%**
STC	13%	92%	26%	-
Zhang *et al.*	-	90%	-	73.7%
MDT	25%	85%	55%	-
MPPCA	30%	77%	-	-
SF+MPPCA	36%	71%	-	-
Adam *et al.*	42%	63%	-	-

[24], Social Force and MPPCA (SF+MPPCA) and Adam *et al.* [22]. The results are displayed in Fig. 10, Fig. 11 and Tab. IV (Please note since most pixel-level ROC Curves for ped2 are not given by literature, we merely plot the ROC curve of our approach and [13]). As shown in Tab. IV, the performance improvement brought by our approach is even greater on ped2 with frame-level EER $< 10\%$ and pixel-level EER $< 20\%$. Such performance gain can be explained by ped2's higher video frame resolution, which can facilitate SL-HOF descriptor to capture optical flow distribution and local motion information in foreground objects.

It should be noted that the performance of our approach is achieved by classic OCSVM instead of more complex event modeling approaches like Sparse Coding, which often calls for higher computation due to involving $l_{2,1}$ and l_1-norm optimization. It demonstrates the effectiveness of our video representation again. Compared with other methods, our approach takes a simple implementation but can achieve comparable or higher detection performance. Examples of detected video abnormal events on UCSD ped1 and ped2 are shown in Fig. 12 and Fig. 13, and we can see different categories of video anomaly can be well detected.

V. Conclusion

In this paper, we propose a simple but efficient approach to detect anomaly from crowded scenes based on SL-HOF

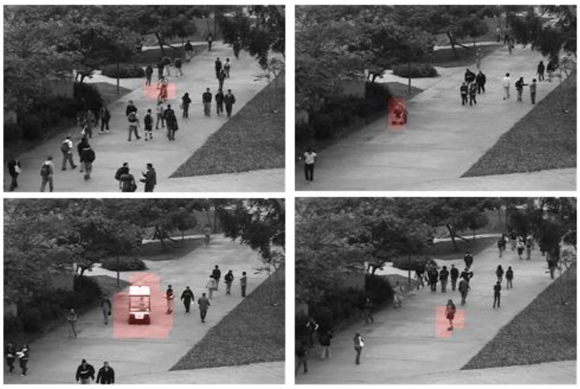

Fig. 12: Detected anomaly in ped1: Biker, wheelchair, vehicle and skater.

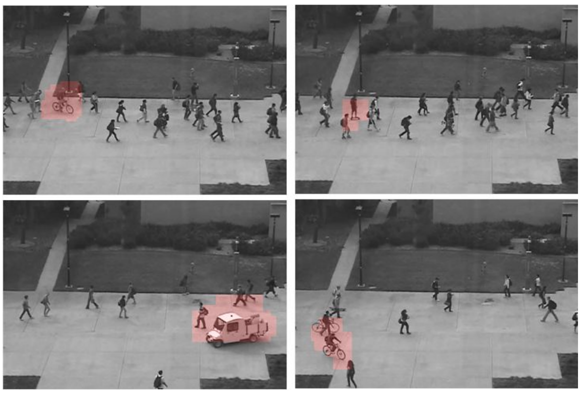

Fig. 13: Detected anomaly in ped2: Biker, skater, vehicle and walking man with bike.

descriptor and foreground classification. The proposed SL-HOF descriptor can capture the spatial distribution of optical flow and local motion information embedded in video foreground objects, which leads to a higher discriminative power than classic video descriptors in video anomaly detection. Foreground classification is proposed to enhance anomaly localization performance by detecting texture anomaly in video frames. The proposed approach yields state-of-the-art results on the challenging UCSD datasets.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Project No. 60970034, 61170287, 61232016).

References

[1] P. C. Chung and C. D. Liu, "A daily behavior enabled hidden markov model for human behavior understanding," *Pattern Recognition*, vol. 41, no. 5, pp. 1572–1580, 2008.

[2] A. Sodemann, M. P. Ross, B. J. Borghetti *et al.*, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1257–1272, 2012.

[3] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[4] T. Zhang, H. Lu, and S. Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1940–1947.

[5] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[6] Z. Fu, W. Hu, and T. Tan, "Similarity based vehicle trajectory clustering and anomaly detection," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, vol. 2. IEEE, 2005, pp. II–602.

[7] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1975–1981.

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

[9] M. Roshtkhari and M. Levine, "Online dominant and anomalous behavior detection in videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2611–2618.

[10] Y. Zhao, Y. Qiao, J. Yang, and N. Kasabov, "Abnormal activity detection using spatio-temporal feature and laplacian sparse representation," in *Neural Information Processing*. Springer, 2015, pp. 410–418.

[11] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1446–1453.

[12] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2013, pp. 2720–2727.

[13] Y. Zhang, H. Lu, L. Zhang, and R. Xiang, "Combining motion and appearance cues for anomaly detection," *Pattern Recognition*, vol. 51, pp. 443–452, 2016.

[14] B. Antic and B. Ommer, "Video parsing for abnormality detection," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2415–2422.

[15] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2909–2917.

[16] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2112–2119.

[17] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008.

[18] J. Wright, A. Ganesh, S. Rao, and M. Yi, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization." *Advances in Neural Information Processing Systems*, vol. 87, no. 4, p. 20:320:56, 2009.

[19] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[20] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2008, pp. 275–1.

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

[22] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[23] M. J. Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1436–1452, 2013.

[24] J. Kim and K. Grauman, "Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2921–2928.