Abnormal Event Detection in Video using Motion and Appearance Information

Neptalí Menejes Palomino[1] and Guillermo Cámara Chávez[2]

[1] Universidad Católica San Pablo, Arequipa, Perú,
neptalimenejes@gmail.com,
[2] Computer Science Department, Federal University of Ouro Preto,
Ouro Preto, MG, Brazil,
gcamarac@gmail.com

Abstract. This paper presents an approach for the detection and localization of abnormal events in pedestrian areas. The goal is to design a model to detect abnormal events in video sequences using motion and appearance information. Motion information is represented through the use of the velocity and acceleration of optical flow and the appearance information is represented by texture and optical flow gradient. Unlike literature methods, our proposed approach provides a general solution to detect both global and local abnormal events. Furthermore, in the detection stage, we propose a classification by local regions. Experimental results on UMN and UCSD datasets confirm that the detection accuracy of our method is comparable to state-of-the-art methods.

Keywords: Abnormal event detection, video analysis, spatiotemporal feature extraction, video surveillance, computer vision.

1 Introduction

In recent years, abnormal event detection in video has attracted more attention in the computer vision research community due to the increased focus on automated surveillance systems to improve security in public places, such as: airports, railway stations, shopping malls, etc. Traditionally, video surveillance systems are monitored by human operators, who alert if there are any suspicious events on the scene. However, human eye is susceptible to distraction or tiredness due to long hours of monitoring. In addition, the relationship between human operators and the large number of cameras is disproportionate, making it more difficult to detect and respond to all anomalous events that occur on the scene. This motivates the need for an automated abnormal event detection framework using computer vision technologies.

An abnormal event has no consistent definition, as it varies according to the context. In general, it is defined as an event which stands out from the normal behavior within a particular context. In this work, we consider the context of pedestrian walkways, where the normal behavior is performed by people walking. Events which involve speed violations and the presence of abnormal objects are considered to be abnormal. We present an approach for the detection and localization of abnormal events in video using motion and appearance information.

Motion information is represented through the use of the velocity and acceleration of optical flow and appearance information is represented by the textures and optical flow gradient. To represent these features we use non-overlapping spatio-temporal patches. Unlike literature methods, our model provides a general solution to detect both, global and local abnormal events. Furthermore, the detection stage presents problems of perspective distortion, this occur due to the objects near the camera appear to be large, while objects away from the camera appear to be small. To address these problems, we propose a classification by local regions.

The remainder of the paper is organized as follows. Section 2 reviews previous work on abnormal event detection. An overview on the proposal is provided in Section 3. Section 4 presents the results of our method on several public datasets. Finally, conclusions are provided in Section 5.

2 Related Works

Depending on the context and the specific application, abnormal event detection generally falls into two categories: trajectory analysis and motion analysis. Trajectory analysis is based on object tracking and typically requires an uncrowded environment to operate. There are some trajectory-based approaches for object tracking [13, 14, 15], these methods can obtain satisfactory results on traffic monitoring, however, fail in crowded scenes, since they cannot get good object trajectories due to the number of objects or events occurring simultaneously. Motion analysis is better suited for crowded scenes by analyzing patterns of movement rather than attempting to distinguish objects. Most of the state-of-the-art methods use local features for abnormal event representation, by considering the spatio-temporal information. For crowded scenes, according to the scale of interest, abnormal event detection can be classified into two [1]: Global Abnormal Event detection (GAE) and Local Abnormal Event detection (LAE). The former focuses on determining whether the scene contains an abnormal event or not, while the latter determines the location where the event is occurring.

To detect GAE, some methods have been proposed. Mehran *et al.* [2] propose a method to detect abnormal crowd behavior by adopting the Social Force Model (SFM), then use Latent Dirichlet Allocation to detect abnormality. Wang *et al.* [7] extract Histograms of Oriented Optical Flow (HOOF) for the detection of GAE, and a Bayesian framework for crowd escape behavior detection [9]. To detect LAE, most approaches extract motion and appearance features from local 2D patches or local 3D cubes. Kratz *et al.* [10] extract spatio-temporal gradient to fit a Gaussian model, and then use HMM to detect abnormal events. Mahadevan *et al.* [11] proposed a detection framework which uses mixture of dynamic textures to represent both motion and appearance features. Raghavendra *et al.* [12] introduce the Particle Swarm Optimization method to optimize the force interaction calculated using SFM. Furthermore, some methods based on optical flow estimation have been proposed. Ryan *et al.* [6] used optical flow vectors to model the motion information and proposed the textures of optical flow features to capture the smoothness of the optical flow field across a region, enabling the detection of objects such as bicycles and vans. Cong *et al.* [8] presented a novel algorithm for abnormal event detection based on the sparse reconstruction cost for multilevel histogram of optical flows.

Recently, deep learning technique has been also used to abnormal event detection. Xu *et al.* [18] proposed a model that trains a Convolutional Neural Network (CNN) using spatio-temporal patches from optical flow images as input. Revathi *et al.* [20] use the deep learning classifier to detect abnormal event. Ravanbakhsh et a*et al.*. [19] introduce a novel Binary Quantization Layer and then propose a temporal CNN pattern measure to represent motion in crowd.

3 Proposed Method

In this section, we present our approach for abnormal event detection, illustrated in Figure 1. We use different sets of features to model the normal events and to detect different anomalies related to speed violations, access violations to restricted areas and the presence of abnormal objects in the scene.

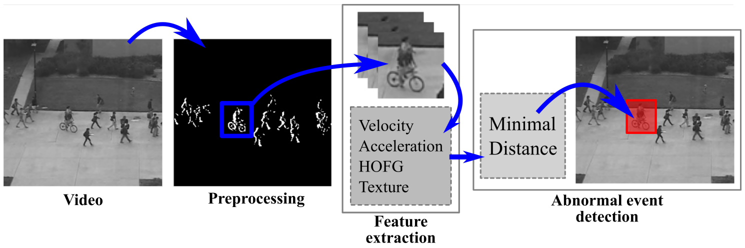


Fig. 1. Workflow of the proposed approach.

3.1 Feature Extraction

In order to restrict the analysis to regions of interest and to filter out distractions (*eg.* waving trees, illumination changes, etc), we perform foreground extraction on each incoming frame. In this work, we calculate the foreground by the binarization of the subtraction of two consecutive frame, generating a foreground mask. In addition, to estimate the optical flow field of each frame, our approach considers the Lucas-Kanade pyramidal optical flow algorithm. It is important to note that only the optical flow of the foreground pixels is calculated.

Our approach divides the video sequence into non-overlapping spatio-temporal local patches of $m \times n \times t$. For each spatio-temporal patch P, we extract features based on motion, such as velocity and acceleration of optical flow, and features based on appearance, such as textures and gradients of optical flow. The extraction process of each of the features is shown below.

Velocity: The optical flow velocity feature is the summation of optical flow vectors inside a spatio-temporal patch [4].

Acceleration: The optical flow acceleration feature extracts information about the temporal variation of optical flow [5]. To model this feature, we calculate the optical flow of each incoming frame and then compute the optical flow magnitude of each foreground pixel, the magnitude is normalized between $[0, 255]$ to create a magnitude image. Again, we compute the optical flow of the magnitude image to calculate the acceleration. Finally, for each spatio-temporal patch the acceleration information is calculated by a summation of the acceleration vectors.

Textures of Optical Flow: The textures of optical flow measure the uniformity of the motion [6]. This feature is computed from the dot product of flow vectors at different offsets p and $p^{'} = p+\delta$, where δ denotes the displacement of the pixel p.

Histogram of Optical Flow Gradient (HOFG): This feature extracts information about the spatial variation of optical flow [5]. To model this feature, we first treat both horizontal and vertical optical flow components as a separate image and then compute the gradients of each image using Sobel operators. For each spatio-temporal patch, we generate a four-bins histogram in each image using a soft binning-based approach and finally concatenate both histograms into a single eight-dimensional feature vector.

3.2 Abnormal Event Detection

In real world, usually normal event samples are predominant, and abnormal event samples have little presence. Because of this, the detection of abnormal events becomes even more difficult. In this paper, to address this problem, we use the classification method proposed in [3].

In [3], the classification stage uses the minimum distance. The main idea of this classification method is to search trained pattern that are similar to the incoming pattern. That is, if the incoming pattern is similar enough to some of the known patterns, then it is considered as a normal pattern. Otherwise, it will be considered as an abnormal event. However, the classification based on the minimum distance presents problems of detection due to some videos present the perspective distortion in the scene. That is, objects near to the camera appear to be large while distant objects appear to be small. This can significantly affect the feature extraction methods as the extracted features will vary according to their depth in the scene. In this paper, to address the problem of perspective distortion and to obtain better results, we propose a classification based on local regions. This classification method divides the scene into local spatial regions according to the depth of the scene, where normal patterns are modeled in each local region.

4 Experimental Results

We test our proposed approach by combining the proposed features (see Section 3) on UMN dataset [16] and UCSD dataset [11]. Our experiments are divided in two parts. The first part shows a comparison of our proposed classification method with 1 and 4 classification regions. Then, the second part compares our results with other methods published in the literature.

Experimental Setup and Datasets: The final feature vector contains two features describing the horizontal and vertical optical flow fields (velocity), two optical acceleration features (horizontal and vertical directions), three features of optical flow textures and eight-dimensional histogram of optical flow gradient components, four components for each vertical and horizontal directions. Therefore, we have a fifteen-dimensional feature vector for each spatio-temporal patch. In addition, we experimentally set a fixed spatio-temporal patch size of $20 \times 20 \times 7$. The criterion used to evaluate abnormal events detection accuracy was based on frame-level and to measure the performance of the proposed method, we have calculated the ROC curve, the area under the curve (AUC) and the equal error rate (EER).

UMN dataset consists of three different scenes of crowded escape events with a 320×240 resolution. The normal events are pedestrians walking randomly and the abnormal events are human spread running at the same time. There are a total of 11 video clips in the dataset. In the training phase the first 300 frames of each video were used to model normal events and the remaining frames were used in the test phase. UCSD dataset includes two sub-datasets, *Ped1* and *Ped2*. The crowd density varies from sparse to very crowded. The training sets are all normal events and the testing set includes abnormal events like cars, bikes, motorcycles and skaters. Ped1 contains 34 video clips for training and 36 video clips for testing with a 158×238 resolution, and Ped2 contains 16 video clips for training and 12 video clips for testing with a 360×240 resolution.

Global Abnormal Event Detection: We use UMN dataset [16] to detect GAE. The results are shown in Fig. 2, where columns (a), (b) and (c) represent the results on UMN dataset (scene 1,2 and 3, respectively). Furthermore, the first row shows the results for 1 region of classification, while the second for 4 regions. The ROC curves of the proposed method are shown in Fig. 3 (a).

In Table 1, we show the quantitative results of our experiment on the UMN dataset, using two ways of classification (1 and 4 regions). In scenarios 1 and 3, our method using the classification with a single region achieves an AUC of 0.9985 and 0.9954, respectively. These results overcome the classification with 4 regions. Meanwhile, in scene 2, the result using 4 regions of classification achieves an AUC of 0.9486, which overcomes the result of the classification using a single region. This happens due to the videos in scene 2 have problems of perspective distortion. Table 2 provides the quantitative comparisons to the state-of-the-art methods. The AUC of our method in scenes 1 and 3 overcomes the state-of-the-art methods, which are 0.998 and 0.995, respectively. However, the AUC in scene 2 is comparable to the results of literature.

Our method	AUC (1 region)	AUC (4 regions)
Scene 1	**0.9985**	0.9962
Scene 2	0.9182	**0.9486**
Scene 3	**0.9954**	0.9951

Table 1. Quantitative results of the proposed method on the UMN dataset.

Methods	AUC
Sparse Scene 1 [8]	0.995
Sparse Scene 2 [8]	**0.975**
Sparse Scene 3 [8]	0.964
Scene 1 [17]	0.936
Scene 2 [17]	0.775
Scene 3 [17]	0.966
Our proposed Scene 1	**0.998**
Our proposed Scene 2	0.948
Our proposed Scene 3	**0.995**

Table 2. Quantitative comparison of the proposed method with the state-of-the-art methods on the UMN dataset.

Local Abnormal Event Detection: We use the UCSD dataset to detect EAL. The results are shown in Fig. 2, where columns (d) and (e) represent the results on Ped1 and Ped2 datasets, respectively. Furthermore, we show the classification of 1 and 4 regions in the first and second row, respectively. The ROC curves of the our method on the UCSD dataset are shown in Fig. 3 (b).

Table 3 shows the quantitative results of our experiment on the UCSD dataset, using two ways of classification (1 and 4 regions). The results on the Ped1 dataset using 4 regions of classification achieves an EER of 29.28% and an AUC of 0.7923, this result overcomes the classification using a single region due to perspective distorsion problem presented in this dataset. On the other hand, the results on the Peds2 dataset using a single region of classification achieves a EER of 07.24% and an AUC of 0.9778, this result overcomes the classification using 4 regions. Table 4 shows quantitative comparison of the proposed method with the state-of-the-art methods on the UCSD dataset. In the Ped1 dataset, our method achieves an EER of 29.2% and an AUC of 0.792, being competitive with most of the reported methods in the literature. On the other hand, on Ped2, our method achieves an EER of 07.2% and an AUC of 0.977, outperforming

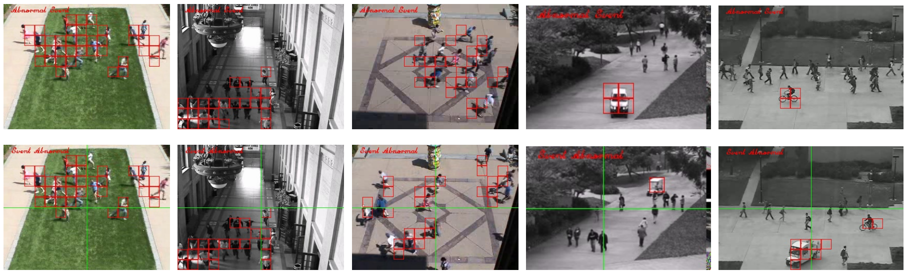
Our	01 region		04 regions	
method	EER	AUC	EER	AUC
Ped1	32.33%	0.737	**29.28%**	**0.792**
Ped2	**07.24%**	**0.977**	07.81%	0.976

Table 3. Quantitative results of the proposed method on the UCSD dataset.

	Ped1		Ped2	
Methods	EER	AUC	EER	AUC
Social force[2]	31.0%	-	42.0%	-
MDT [11]	25.0%	-	25.0%	-
Reddy [4]	22.5%	-	20.0%	-
Textures [6]	23.1%	0.838	12.7%	0.939
HOFM [3]	33.3%	0.715	19.0%	0.899
CNN [18]	16.0%	0.921	17.0%	0.908
TCNN [19]	**08.0%**	**0.957**	18.0%	0.884
Our proposed	29.2%	0.792	**07.2%**	**0.977**

Table 4. Quantitative comparison of the proposed method with the state-of-the-art methods on the UCSD dataset.

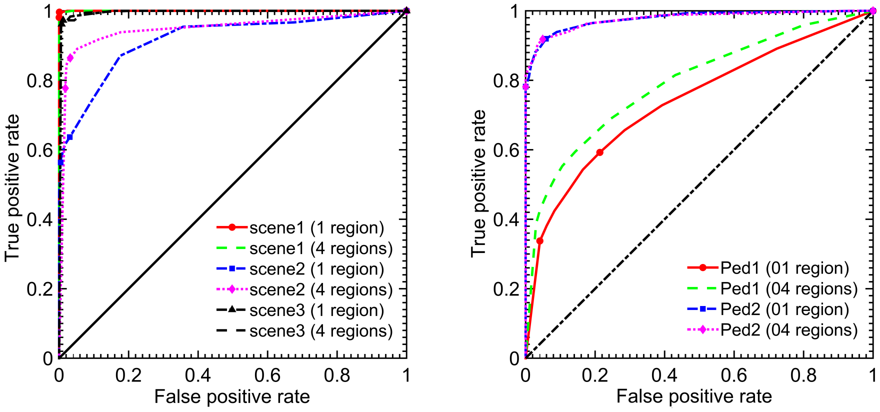
all reported results. It is important to emphasize, that state-of-the-art methods that use deep learning techniques [18, 19] achieve better results on Ped1 dataset. However, our proposal overcomes the results of all methods on Ped2 dataset.



(a) UMN scene1 (b) UMN scene2 (c) UMN scene3 (d) UCSD ped1 (e) UCSD ped2

Fig. 2. Results of the proposed method on the UMN and UCSD dataset.

In our experiments we observed that the results on video sequences that don't have problems of perspective distortion, the performance of our method overcomes all state-of-the-art methods including techniques that use deep learning. This is because our proposed extracts and combines information of motion and appearance. However, our results fall down when videos have problems of perspective distortion. To address this problem we propose a classification by local regions, which improves the performance of our method, as can be observed in Fig. 3 (a)(b).



(a) ROC curves on UMN dataset (b) ROC curves on UCSD dataset

Fig. 3. The ROC curves for the detection of GAE on the UMN dataset.

5 Conclusions

In this paper, we propose a new approach based in the motion and appearance information for abnormal event detection (GAE and LAE) in crowded scenes. Motion features are suitable for detecting abnormal events with high speed motion. However, some abnormal events have a normal speed, to address this type of events, we introduce the use of appearance features. Furthermore, we propose a new classification method based on local regions to address the problems of perspective distortion in videos.

We evaluated the performance of our proposed method on the UCSD and UMN datasets. On the UCSD Ped1 dataset and scene 2 of UMN dataset, our results are comparable with the literature methods. On the other hand, our results on the UCSD Ped2 dataset achieves an EER of 07.2% and an AUC of 0.977, and on the UMN dataset achieves an AUC Of 0.998 and 0.995 in the scenes 1 and 3, respectively. According to these results, our method achieved better results and overcome the state-of-the-art results.

Acknowledgment

This work was supported by grant 011-2013-FONDECYT (Master Program) from the Nacional Council for Science, Technology and Technological Innovation (CONCYTEC-PERU).

References

[1] Li, T., Chang, H., Wang, M., Ni, B., Hong, R., and Yan, S. (2015). Crowded scene analysis: A survey. In TCSVT 2015, pp. 367-386. IEEE.

[2] Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In CVPR 2009. (pp. 935-942). IEEE.

[3] Colque, R. V. H. M., Júnior, C. A. C., and Schwartz, W. R. (2015). Histograms of optical flow orientation and magnitude to detect anomalous events in videos. In SIBGRAPI 2015, (pp. 126-133). IEEE.

[4] Reddy, V., Sanderson, C., and Lovell, B. C. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In CVPRW 2011, (pp. 55-61). IEEE.

[5] Nallaivarothayan, H., Fookes, C., Denman, S., and Sridharan, S. (2014). An MRF based abnormal event detection approach using motion and appearance features. In AVSS 2014, (pp. 343-348). IEEE.

[6] Ryan, D., Denman, S., Fookes, C., and Sridharan, S. (2011). Textures of optical flow for real-time anomaly detection in crowds. In AVSS 2011, (pp. 230-235).

[7] Wang, T., and Snoussi, H. (2013). Histograms of optical flow orientation for abnormal events detection. In PETS 2013, (pp. 45-52). IEEE.

[8] Cong, Y., Yuan, J., and Liu, J. (2013). Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition, 46(7), 1851-1864.

[9] Wu, S., Wong, H.-S., et al. (2014). A bayesian model for crowd escape behavior detection. In CSVT 2014, 24(1):85-98. IEEE.

[10] Kratz, L., and Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In CVPR. (pp. 1446-1453).

[11] Mahadevan, V., Li, W., et al. (2010). Anomaly detection in crowded scenes. In CVPR 2010, pp. 1975-1981. IEEE.

[12] Raghavendra, R., Del Bue, A., Cristani, M., and Murino, V. (2011). Optimizing interaction force for global anomaly detection in crowded scenes. In ICCV 2011, (pp. 136-143). IEEE.

[13] Guler, S., and Farrow, M. K. (2006). Abandoned object detection in crowded places. In Proc. of PETS (pp. 18-23). Citeseer.

[14] Gul, S., Meyer, J. T., Hellge, C., Schierl, T., and Samek, W. (2016). Hybrid video object tracking in H. 265/HEVC video streams. In MMSP, (pp. 1-5). IEEE.

[15] Ngo, D. V., Do, N. T., and Nguyen, L. A. T. (2016). Anomaly detection in video surveillance: A novel approach based on sub-trajectory. In ICEIC, (pp. 1-4). IEEE.

[16] Unusual crowd activity dataset of University of Minnesota, from http://mha.cs.umn.edu/movies/crowdactivity-all.avi.

[17] Shi, Y., Gao, Y., and Wang, R. (2010). Real-time abnormal event detection in complicated scenes. In ICPR 2010, (pp. 3653-3656). IEEE.

[18] Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection.

[19] Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., and Sebe, N. (2016). Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection.

[20] Revathi, A. R., and Kumar, D. (2016). An efficient system for anomaly detection using deep learning classifier. Signal, Image and Video Processing, 1-9.