Violent Event Detection: An Approach Using Fusion GHOG-GIST Descriptor

B. H. Lohithashva, V. N. Manjunath Aradhya, and D. S. Guru

Abstract In this paper, we propose violent event detection in video using fusion of global histograms of oriented gradients (GHOG) and GIST feature descriptors. The significant features are extracted from GHOG, GIST, and GHOG + GIST. Finally, significant features are used with the support vector machine (SVM) classifier for the detection of a violent event. The proposed feature descriptor method used Hockey Fight and Violent-Flows dataset for the experimentation, and empirical results show that the proposed GHOG + GIST fusion feature descriptor is effective and efficient than other state of the art techniques.

Keywords Violent event · Texture features · GHOG · Descriptor · GIST · SVM

1 Introduction

Violent event detection is one of the most interesting and difficult research problems in action recognition, and it is a dynamic research in machine learning and computer vision. Video surveillance can be illustrated as the task of studying videos to detect unusual events, and it is applicable in both private and public sectors like in traffic control, health care, airports, museum, railway stations, etc. Due to its widespread applications, video surveillance has become the medium to understand human behavior and helps in providing reliable security to the people [1]. The violent event detection is referred to as the pattern to be detected from the video that is not same as normal behavior. The actions which take place are not normal such as fraud detection, slip and fall detection, intrusion and loitering detection, event in a sensor network, and much more.

B. H. Lohithashva (✉) · V. N. Manjunath Aradhya
Department of Computer Applications, JSS Science and Technology University (Sri Jayachamarajendra College of Engineering), Mysuru, Karnataka, India
e-mail: lohi.bh@gmail.com

D. S. Guru
Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka, India

V. Komanapalli et al. (eds.), *Advances in Automation, Signal Processing, Instrumentation, and Control*, Lecture Notes in Electrical Engineering 700,
https://doi.org/10.1007/978-981-15-8221-9_82

In this paper, we propose the fusion of shape and texture-based features such as GHOG/GIST descriptor to detect the violent event in the publically available Hockey Fight and Violent-Flows standard benchmark datasets. Our feature extraction technique is demonstrated on the combination of the histograms of orientation, scale, and orientation by convolving it with Gabor filters in a video scene. The additional information to the feature descriptor, to extract the histograms of orientation magnitude, and spatial envelope spectrum of texture features capture the motion correctly. We use the support vector machine (SVM) classifier to distinguish between violent and non-violent events. To increase the accuracy rate, we accomplished the space-time post-processing technique. The proposed GHOG + GIST descriptor being efficient outperforms the other states of the art methods as demonstrated through experimentation. The rest of this paper is planned as follows. Section 2, we present previous works on violent event detection and contribution. The proposed shape and texture-based feature descriptors are explained in Sect. 3. Experimentation results are confronted and are discussed in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Previous Works and Contribution

Nowadays, many researchers focused on violent events and have proposed different feature descriptor techniques [2, 3], feature selection techniques [4, 5], and classifiers [6, 7] to detect violent events [8, 9]. Dalal and Triggs [10] for the first time introduced histograms of oriented gradients (HOG) descriptor for human detection and later on, modified HOG was used for abnormal activity detection. Lohithashva et al. [1] introduced the technique to detect the usual and unusual event of the crowd using a holistic feature descriptor from the histogram. Gao et al. [11] introduced violent event detection using oriented ViF (OViF) descriptor, which is an extension of ViF descriptor. OViF descriptor extracts reliable features based on the motion gradient magnitude. Compared to ViF descriptor, the OViF descriptor is efficient in both crowded and un-crowded behavior. Xia et al. [12] introduced violent event detection using deep learning technique where the method mainly focused on spatial and temporal information in the video to detect the violent and non-violent events. Oliva and Torralba's [13] have first introduced the GIST feature descriptor for scene categorization. Later on, many researchers have used for object detection and autonomous car driving. Even though many researchers have been using different feature descriptors to detect the violent event, still there are some limitations that make it challenging. Actually, a sudden change in the background, extraction of pertinent information from little changes in the object appearance is difficult to identify. In this work, we have used GHOG, GIST, and GHOG + GIST features fusion to SVM classier to distinguish between violent and non-violent events. GHOG is an effective feature descriptor for shape or edge information of an object, but it is poorly performed if there is a complex background with noise. Merely, the GIST feature descriptor works well even if it is a complex background and loft varied illumination. Consequently, we fuse GHOG and GIST descriptor to improve the detection of violent event.

The contributions of the paper are as follows:

The traditional HOG feature descriptor was initially used for human detection. The feature is generated for each frame of size 34596. It is the first of the kind in literature, and the GHOG and GIST feature descriptors are used for violent event detection and we have downsized the features of GHOG to 72 dimensions and GIST to 288 dimensions. Then, the fusion of GHOG + GIST features vector size of each frame to 360 dimensions, which nevertheless produces a good result.

GIST descriptor was used for scene categorization and for the first time we have applied the GIST descriptor to detection of violent events.

The combined space-time post-processing method is used to improve the accuracy of the violent event detection.

3 Proposed Methodology

We fuse the GHOG and GIST feature descriptors to detect violent events. The architecture of the proposed method is as shown in Fig. 1. The proposed GHOG-GIST features are extracted in 3×3 window, the succeeding section, we will give an overview of the GHOG and GIST feature descriptors and a detailed implementation aspect of the proposed method.

3.1 GHOG Feature Descriptor

GHOG is an effective global feature descriptor that can be used to extract shape or edge information. The gradients of a frame are determined as the directional change of the frame intensity or color. Merely, it can be prevailed horizontally u_x by convolving the frame with the one-dimensional gradient templet operator $(-1 \;\; 0 \;\; 1)$ and vertically v_y by convolving the frame with $(-1 \;\; 0 \;\; 1)^T$. We calculate the magnitude of the gradient M and orientation θ in Eqs. (1) and (2), respectively.

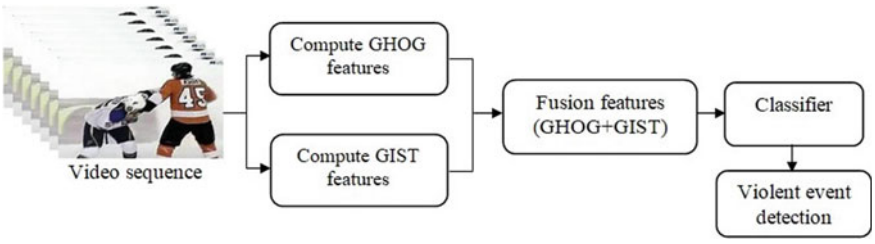

Fig. 1 Architecture of the proposed method

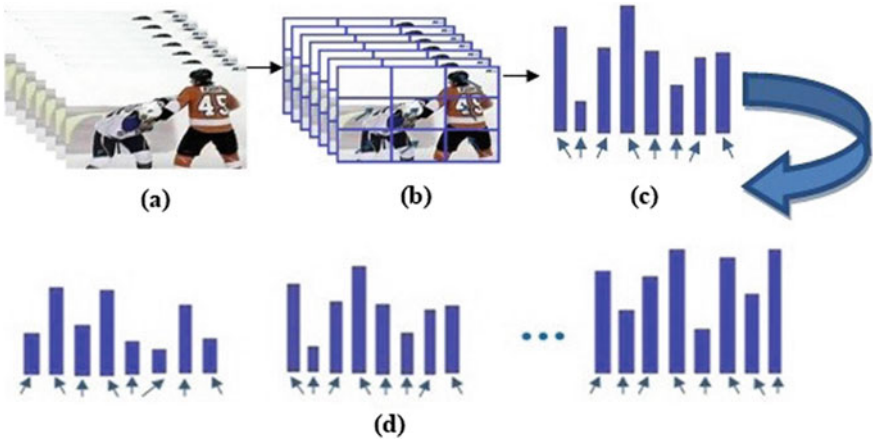

Fig. 2 GHOG feature extraction procedure; **a** Input video sequence, **b** Construction of histograms of oriented gradients in the 3 × 3 window, **c** Distribution oriented is normalized to obtain an 8 bin histograms, d GHOG feature is computed

$$M = \sqrt{u_x^2 + v_y^2} \tag{1}$$

$$\theta = \tan^{-1}\left(\frac{v_y}{u_x}\right) \tag{2}$$

The proposed GHOG feature descriptor is demonstrated in Fig. 2. The sequence of the input frames is divided into 3 × 3 square window and arranged in a number of successive frames is established about the window demonstrate in Fig. 2.

The gradients magnitude and their orientation angles for each pixel in the regions of the cells are calculated. Each cell is represented as $1 \times bin_n$ row vector. Each cell in the window gets a different row vector which is combined into a long row vector. The histograms of gradients magnitude M and orientation angles θ for each pixel in the cells are calculated. Each cell is discretized into angular bins according to the orientation interval. The histograms of oriented gradients of the 8 cells in the window are computed for the interval [0°–180°] with an 8-orientation binning size. The feature vectors of the window regions should be normalized to L_2 norm. The overall histograms of oriented gradients for the entire frame are constructed by the concatenation of the whole histograms extracted from the window regions in one feature vector representing the GHOG features of the frame, and 72 (3 × 3 × 8) dimensions of GHOG features for each frame are obtained.

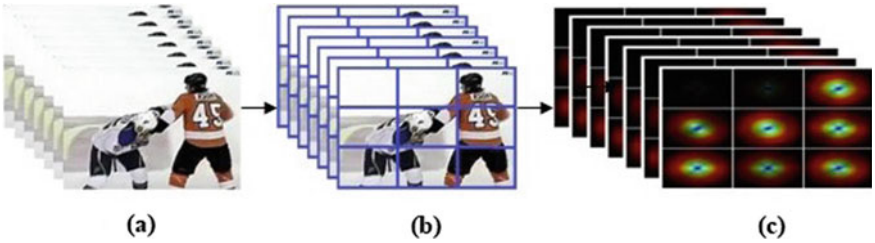

Fig. 3 GIST feature extraction procedure: **a** Input video Sequence **b** Construction of 3 × 3 window **c** Construction of spatial structural information of orientations and scales of GIST features is computed

3.2 GIST Feature Descriptor

The GIST feature descriptor is a global texture-based feature extraction technique which used extracting spatial structural information of frames without segmentation of frames. Initially, divide frames into 3 × 3 window, the sequence of frames are converted into gray-scale frames, and it is handled by a whitening filter; the dominant spatial structural information of the frames can be saved. The filters are arranged at 4 scales (σ) and 8 orientations (θ) leading in a series of 32 features. To accomplish the gray-scale frames that are convolved with the series of Gabor filters is gaussian kernel function [14, 15], as shown in Eqs. (3)–(5). For each cell in the window, the average intensity of the cell is computed. Finally, the extracted features are concatenated and resulting in a 288-dimensional GIST feature vectors for each frame.

$$G(p, q, \sigma, \theta) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{p'^2}{2\sigma^2} + \frac{q'^2}{2\sigma^2}\right)} e^{j\frac{\pi p'}{\sigma}} \tag{3}$$

$$p' = p\cos\theta + q\sin\theta \tag{4}$$

$$q' = -p\sin\theta + q\cos\theta \tag{5}$$

where p and q represents x- and y-axis of the frame, σ represents scale, and θ represents orientation. A set of spatial energy envelope spectrum with different scale and orientation is used to extract prominent features as shown in Fig. 3.

3.3 GHOG-GIST Descriptor

In this section, we described violent event detection based on the fusion of GHOG and GIST features descriptors in both crowded and uncrowned video scenes. First, the GHOG feature descriptor is used for feature extraction based on the motion of the

gradient information and we get 72 features for each frame. Second, we use the GIST descriptor for feature extraction based on the scale and orientation by convolving it with Gabor filters resulting in 288 features which is for each frame. GHOG and GIST features are fused using the features fusion scheme, and 360 features for each frame are obtained.

Finally, features are used with the SVM classifier for the classification of violent and non-violent events. The accomplished illustration of violent event detection using the proposed features fusion descriptor is as shown in Fig. 1.

3.4 Classifier

SVM classifier is a supervised learning model that can be used to analyze data for classification, first time introduced by Vladimir Vapnik [16]. The SVM classifier considers non-parametric functions when the feature vectors do not separate linearly. Because it relies on kernel functions and hyper-plane, it pretends like a decision boundary. Hyper-plane segregates input feature vector into two or more classes in an n-dimensional space, where n represents the number of input features. In this research work, we have used coarse gaussian kernel function to classify violent and non-violent event.

3.5 Post-processing

The process of post-processing helps to improve the accuracy of the prediction and minimize the false positive rate. Wang et al. [17] for the very first time introduced temporal post-processing technique and Vikas et al. [18] used space-time post-processing method. In this paper, we have used the post-processing technique proposed by [19]. We employ space-time post-processing proposed by taking the number of frames to 30 instead of 10 for the detection of frame-level.

4 Experimental Results and Discussion

In this section, we demonstrate the experimental results carried to analyze the proposed method of violent event detection on two standard benchmark datasets [11]. Subsequently, the experimental setting is elaborated. Finally, the obtained results are compared with that of other states of the art feature descriptors.

Fig. 4 **a** Hockey Fight dataset sample frames, **b** Violent-Flows dataset sample frames

4.1 Violent Dataset

In this work, we have used Hockey Fight dataset and Violent-Flow dataset to being onto the significant of our Fusion features descriptor for both crowded and uncrowded scenes for violent event detection. **Hockey Fight dataset** is particularly employed to evaluate the detection of violent event systems [11]. It comprises of 1000 video clips (500 violent and 500 non-violent) of National Hockey League (NHL) action videos. Each video exists fights between two or elite players and it is a non-crowded violent dataset. Figure 4a shows various samples of frames consisting of wrestle and no wrestle scenes from the dataset. **Violent-Flows dataset** is used to evaluate crowded violent event detection [11]. This dataset comprises of 246 action videos (123 violent and 123 non-violent from different scenes). All the violent videos are downloaded from the web and each video is an average of 3.60 s and is underneath unrestrained in the furious conditions. Figure 4b depicts samples consisting of both violence and non-violence scenes from the Violent-Flows dataset.

4.2 Experimentation Setting

In an experimentation, HOG [20], HOF [20], HNF [20], LTP [21], ViF [11], OViF [11], ViF + OViF [11], and DiMOLIF [22] eight baseline state of the art methods are compared with our proposed GHOG + GIST descriptor. In this section, we have demonstrated our proposed feature descriptor performance. In experimentation, we adopt the five-fold cross-validation technique as followed in [11, 22]. There are five folders, each folder are the same ratios between both violent and non-violent videos. Every time we have selected four folders for training, and the left one is used for testing. In Hockey Fight dataset, we have used 32 thousand frames for training purpose and about 8 thousand frames for testing purpose, similarly, in the Violent-Flows fight have used 16 thousand frames for training purpose and about 4 thousand frames for testing purpose. We do this process five times. Finally, we have taken an average of all the results. After getting GHOG, GIST, and GHOG + GIST

features, we fed to SVM classifier to evaluate the performance of the classifier. Our proposed method gives good results than state of the art methods. For the frame-level measurement, we have employed the area under the receiver operating characteristic (AUC) to evaluate the detection and classification accuracy of an algorithm. In the receiver operating characteristic (ROC) curve, correctly detected events are shown on the *y*-axis and false detection of an event is shown on the *x*-axis.

4.3 Results and Analysis

Experimentation is carried on two standard benchmark datasets independently. In the Hockey Fight produces dataset video comprises uncrowned scenarios. Our proposed feature extraction technique is a substantial result. The ROC curves of SVM classifier using GHOG, GIST, GHOG + GIST descriptors are illustrated in Fig. 5. Table 1 provides the details of the SVM classifier performance using GHOG, GIST, GHOG + GIST features, and it can be seen that GHOG + GIST gives the highest accuracy than GHOG and GIST descriptors. Accuracy of 91.18, AUC of 93.45 on the Hockey Fight dataset and accuracy of 88.86, AUC of 92.00 on the Violent-Flows dataset are obtained by the proposed method. The features fusion descriptor has outperformed all other state of the art methods as shown in Table 2. Certainly, our proposed features

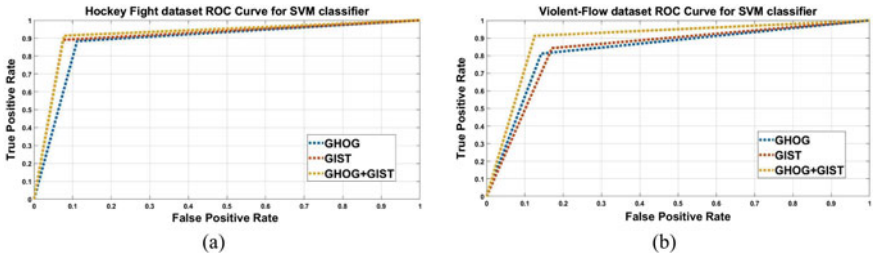

Fig. 5 **a** Hockey Fight dataset ROC Curves for SVM classifier using GHOG, GIST, and GHOG + GIST features descriptor. **b** Violent-Flows dataset ROC curves for SVM classifier using GHOG, GIST, and GHOG + GIST features descriptor

Table 1 The comparison of **GHOG descriptor**, **GIST descriptor,** and **GHOG + GIST descriptor** accuracy with standard deviation (SD) and AUC of SVM classifier is demonstrated in percentage using Hockey Fight dataset and Violent-Flows dataset

Feature descriptor	Hockey Fight dataset		Violent-Flows dataset	
	Accuracy (±SD)	AUC	Accuracy (±SD)	AUC
GHOG	87.56 ± 3.86	89.02	81.85 ± 6.73	83.64
GIST	90.99 ±2.80	91.72	83.54 ± 2.94	84.61
GHOG + GIST	**91.18 ±2.95**	**93.45**	**88.86 ± 5.12**	**92.00**

Table 2 Performance comparisons of all other descriptors accuracy with standard deviation (SD) and AUC are presented in percentage using Hockey Fight and Violent-Flows dataset

Method	Hockey Fight dataset		Violent-Flows dataset	
	Accuracy (±SD)	AUC	Accuracy (±SD)	AUC
HOG	87.8	–	57.43 ± 0.37	61.82
HOF	83.5	–	58.53 ± 0.32	57.60
HNF	87.5	–	56.52 ± 0.31	59.94
LTP	71.90 ± 0.49	–	71.53 ± 0.17	79.86
ViF	81.60 ± 0.22	88.01	81.20 ± 1.79	88.04
OViF	84.20 ± 3.33	90.32	76.80 ± 3.90	80.47
ViF + OViF	86.30 ± 1.57	91.93	86.00 ± 1.41	91.82
DiMOLIF	88.6 ± 1.2	93.23	85.83 ± 4.2	89.25
GHOG + GIST	**91.18 ±2.95**	**93.45**	**88.86 ± 5.12**	**92.00**

fusion descriptor result shows that it is efficient for detection of crowded and un-crowded violent event scenario.

5 Conclusion

In this work, we have proposed an effective GHOG + GIST fusion features descriptor which is capable of a detecting violent and non-violent event. The descriptor is mainly based on the motion of an object gradient and texture or spatial energy envelope spectrum of square patches of the frame at a range of orientations and scales in both crowded and un-crowded scenes. GHOG, GIST, and GHOG + GIST features fusion are fed to SVM classifier to distinguish between violent and non-violent events. Besides, in an experiment, we have used two standard benchmark datasets and the experimental result shows that our proposed GHOG + GIST features fusion descriptor is robust and produces better than state of the art methods. In future, we intend to continue the proposed feature descriptor to more composite video applications with different classifiers.

Acknowledgements The first author is grateful to UGC under RGNF (Rajiv Gandhi National Fellowship) for supporting financially, Letter No. F1-17.1/2014-15/RGNF-2014-15-SC-KAR-73791/(SA-III/Website), JSS Science and Technology University, Mysuru, Karnataka, India.

References

1. Lohithashva BH, Manjunath Aradhya VN, Basavaraju HT, Harish BS (2019) Unusual crowd event detection: an approach using probabilistic neural network. In: Information systems design and intelligent applications. Springer, Berlin, pp 533–542

2. Yan M, Meng J, Zhou C, Tu Z, Tan YP, Yuan J (2020) Detecting spatio-temporal irregularities in videos via a 3D convolutional autoencoder. J Vis Commun Image Representation, pp 1–8
3. Duan Y, Peng T, Qi X (2020) Active contour model based on LIF model and optimal DoG operator energy for image segmentation. Optik 202:1–16
4. Ahmed M, Manjunath Aradhya VN (2016) A study of sub-pattern approach in 2D shape recognition using the PCA and ridgelet PCA. Int J Rough Sets Data Anal (IJRSDA) 3(2):10–31
5. Hanumantharaju MC, Ravishankar M, Rameshbabu DR, Manjunath Aradhya VN (2014) A new framework for Retinex based color image enhancement using particle swarm optimization. arXiv preprint arXiv: 1409.4046
6. Mahantesh K, Manjunath Aradhya VN, Niranjan S (2014) A study of subspace mixture models with different classifiers for very large object classification. In: 2014 international conference on advances in computing, communications and informatics (ICACCI). IEEE, New York, pp 540–544
7. Kumar HM, Harish BS, Kumar SV, Aradhya VNM (2018) Classification of sentiments in short-text: an approach using mSMTP measure. In: Proceedings of the 2nd international conference on machine learning and soft computing. ACM, pp 145–150
8. Dhiman C, Vishwakarma DK (2019) A review of state-of-the-art techniques for abnormal human activity recognition. Eng Appl Artif Intell 77:21–45
9. Manjunath Aradhya VN, Basavaraju HT, Guru DS (2019) Decade research on text detection in images/videos: a review. Evol Intell 1–27
10. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, (2005). CVPR 2005, vol 1. IEEE, New York, pp 886–893
11. Gao Y, Liu H, Sun X, Wang C, Liu Y (2016) Violence detection using oriented violent flows. Image Vis Comput 48:37–41
12. Xia Q, Zhang P, Wang J, Tian M, Fei C (2018) Real time violence detection based on deep spatio-temporal features. In: Chinese conference on biometric recognition. Springer, Cham, pp 157–165
13. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vision 42(3):145–175
14. Bovik AC, Clark M, Geisler WS (1990) Multichannel texture analysis using localized spatial filters. IEEE Trans Pattern Anal Mach Intell 1:55–73
15. Manjunath Aradhya VN, Pavithra MS (2016) A comprehensive of transforms, Gabor Filter and K-means for text detection in images and video. In: Applied computing and informatics (Elsevier), vol 12, pp 109–116
16. Vapnik V, Chapelle O (2000) Bounds on error expectation for support vector machines. Neural Comput 12(9):2013–2036
17. Wang T, Snoussi H (2014) Detection of abnormal visual events via global optical flow orientation histogram. IEEE Trans Inf Forensics Secur 9(6):988–998
18. Reddy V, Sanderson C, Lovell BC (2011) Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: CVPR 2011 workshops. IEEE, New York, pp 55–61
19. Patil N, Biswas PK (2017) Detection of global abnormal events in crowded scenes. In: 2017 twenty-third national conference on communications (NCC). IEEE, New York, pp 1–6
20. Yeffet L, Wolf L (2009) Local trinary patterns for human action recognition. In: 2009 IEEE 12th international conference on computer vision. IEEE, New York, pp 492–497
21. Laptev I, Lindeberg T (2003) Interest point detection and scale selection in space-time. In: International conference on scale-space theories in computer vision. Springer, Berlin, pp 372–387
22. Mabrouk AB, Zagrouba E (2017) Spatio-temporal feature using optical flow based distribution for violence detection. Pattern Recogn Lett 92:62–67