Detection of Violent Events in Video Sequences based on Census Transform Histogram

Felipe de Souza*†, Helio Pedrini*

*Institute of Computing, University of Campinas (UNICAMP)
Campinas, SP, Brazil, 13083-852
†Eldorado Research Institute
Campinas, SP, Brazil, 13083-898

***Abstract*—Video surveillance systems have enabled the monitoring of complex events in several places, such as airports, banks, streets, schools, industries, among others. Due to the massive amount of multimedia data acquired by video cameras, traditional visual inspection by human operators is a very tedious and time consuming task, whose performance is affected by fatigue and stress. A challenge is to develop intelligent video systems capable of automatically analyzing long sequences of videos from a large number of cameras. This work describes and evaluates the use of CENTRIST-based features to identify violence context from video scenes. Experimental results demonstrate the effectiveness of our method when applied to two public benchmarks, Violent Flows [1] and Hockey Fights [2] datasets.**

***Index Terms*—Video Analysis; Violent Detection; Surveillance Systems; Anomalous Events.**

I. Introduction

Due to the advances in digital video technology, large volumes of data have been acquired, stored and transmitted, which makes it impracticable to verify their content by human operators. Such demand promotes the research and development of automatic video analysis systems to deal with massive amounts of videos in a fast and scalable way.

The recognition of human actions [3], [4], [5], [6] through video processing is useful for diverse domains, such as surveillance, intelligent homes, health monitoring, crime prevention, human-computer interface, among others. In particular, detection of violent scenes has received substantial interest in the last years.

The task of analyzing and identifying abnormal patterns in video sequences depends on several factors, such as background, occlusion, camera resolution, amount of people present in the scene, domain context, among others.

In this work, we propose and analyze a novel method for detecting violent events in video sequences. It consists in five main stages. Initially, the video sequences are preprocessed in order to improve the perception about objects in the scene, where operations to reduce the influence of lighting changes are applied. Then, the CENsus TRansform hISTogram (CENTRIST) descriptor [7] is used to extract a set of features from the video frames. Dimensionality of the extracted features is also tested in order to reduce redundant or noisy information while maintaining the most representative characteristics. Finally, the video frames are classified as violent or non-violent.

Experiments are conducted on two public data sets. The results obtained with the proposed method are compared to other approaches available in the literature, achieving competitive recognition accuracy rates.

The remainder of the paper is organized as follows. Sections II and III briefly describe some important concepts and works, respectively, related to the topic under investigation. Section IV presents the methodology proposed in this work, describing the preprocessing, the feature extraction, the feature reduction, as well as the classification process. Section V describes and analyzes the experimental results. Section VI concludes the paper with final remarks and directions for future work.

II. Related Concepts

In this section, we briefly review some relevant concepts related to the topic under investigation.

A. Violent Actions

Research in the action recognition field [3], [4], [5], [6], [8], [9], [10], [11] has advanced significantly over the last decades. Early works were conducted on datasets containing simple actions performed by a single individual. Recent research focuses on more realistic scenarios, in particular for crowded scenes [12], [13], [14].

Two benchmarks, named as Violent Flows [1] and Hockey Fights [2], have been largely used by the community to detect violent events in videos. Several methods have been applied to these datasets by exploring different visual features, such as texture, color, shape and motion [1], [15], [16], [17].

B. Census Transform Histogram

The holistic descriptor known as CENTRIST (CENsus TRansform hISTogram) [7] is employed in our work to extract features from the video frames. This feature descriptor was chosen due to its properties for encoding structural information while suppressing detailed textural information. It models the distribution of local structures and geometrical information through spatial descriptors.

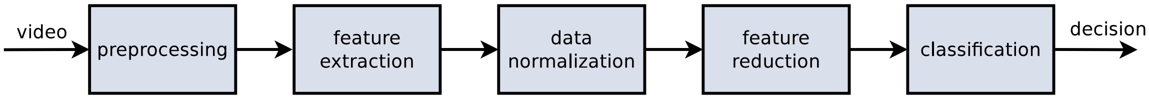


Fig. 1. Main stages of the proposed methodology.

CENTRIST compares the intensity value of the center pixel with its eight neighbors, as shown in Equation 1. If the center pixel intensity value is higher than or equal to one of its neighbors, value 1 is assigned to a bitmap at the corresponding location; otherwise, value 0 is assigned to the bit.

$$\begin{array}{|c|c|c|}\hline 32 & 64 & 96 \\ \hline 32 & 64 & 96 \\ \hline 32 & 32 & 96 \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|}\hline 1 & 1 & 0 \\ \hline 1 & & 0 \\ \hline 1 & 1 & 0 \\ \hline \end{array} \Rightarrow (11010110)_2 \Rightarrow (214)_{10} \quad (1)$$

All Census Transform values are then concatenated and converted into unsigned 8-bit integers represented in the interval between 0 and 255. A histogram with 256 bins is generated to represent the appearance frequency of the Census Transform. The resulting histogram is normalized to the range $[0, 1]$ in order to allow the evaluation of images with different dimensions.

III. Related Work

A method for violent scene identification was proposed by Nam et al. [18], where a spatio-temporal dynamic activity, an audio-visual flame detector, and a blood detector were employed as feature descriptors. Giannakopoulos et al. [19] described an approach to identifying violent videos on video sharing sites by fusing 7 audio features with 1 visual feature.

Nievas et al. [2] employed two spatio-temporal descriptors, Space-Time Interest Points (STIP) [20] and Motion SIFT (MoSIFT) [21], associated with bag-of-words to discriminate local image features, providing a compact representation for the patterns. Support Vector Machines (SVM) with different kernels were explored in the experiments.

Hassner et al. [1] proposed a representation, called Violent Flows (ViF) descriptor, for real-time crowd violence detection. Magnitudes of the optical flow are used to model the frequencies of the ViF words as bag-of-features. A linear SVM was employed to classify the video sequences.

Gao et al. [16] described a feature descriptor, called Oriented Violent Flows (OViF), which explores information of motion magnitude changes in statistical motion orientations. The combination of features using AdaBoost and linear SVM achieved high accuracy rate on the Violent Flows benchmark.

Marsden et al. [15] presented a combination of crowd collectiveness and crowd conflict to model the interaction of objects in crowded scenes. Gaussian Mixture Model (GMM) [22], [23] and SVM were investigated, achieving real-time processing performance on the evaluated datasets.

Zhang et al. [24], [25] described two WLD-based violence detection approaches, named Motion Weber Local Descriptor (MoWLD) and Motion Improved WLD (MoIWLD). A sparse representation-based classification (SRC) was also proposed to minimize the decision error by controlling the reconstruction of coding coefficients. The performance of the methods was evaluated on three datasets.

IV. Methodology

This section describes the proposed methodology for violent event detection. Figure 1 illustrates the main stages of our method, which are detailed in the following sections.

A. *Preprocessing*

Each individual video frame is initially preprocessed in order to make it more suitable for further processing. Different procedures were applied to evaluate their influence in the classification process, including (i) a 3 × 3 Gaussian kernel to reduce noise effect; (ii) a histogram equalization to distribute pixel intensities to a larger contrast range; (iii) a background subtraction using Mixture of Gaussians (MoG) [26] in order to avoid objects not related to the actors of the scene.

Additionally, the video frames are evaluated at multiple scales, where their dimensions were reduced or increased by a constant factor.

B. *Feature Extraction*

The CENTRIST descriptor [7] was chosen due to its properties for encoding structural information while suppressing detailed textural information. It models the distribution of local structures and geometrical information through spatial similar descriptor vectors. Furthermore, HoG descriptor was employed along with CENTRIST to evaluate structural characteristics of the image. Both descriptors were then concatenated to generate a higher dimension descriptor.

The descriptor was employed to extract feature vectors from the video frames in two different strategies: (i) the descriptor was used to evaluate the whole content of each frame; (ii) a grid with specific dimension was used to split each frame into blocks.

We used a sliding window, as shown in Figure 2, with specific size (64×64, 72×72, 96×96 and 128×128 pixels), that traverses each frame by steps with half of the specified block size, such that each block can be individually evaluated. The level of optical flow in each block is evaluated using a threshold value to determine whether the block is relevant or not.

In addition to these feature extraction methods, we also explored some other strategies:

(i) a multi-resolution processing (Figure 3(a)), where the features are extracted from the current frame at different resolutions [27];

(ii) a spatio-temporal processing (Figure 3(b)), where the descriptor is composed from sequential blocks;

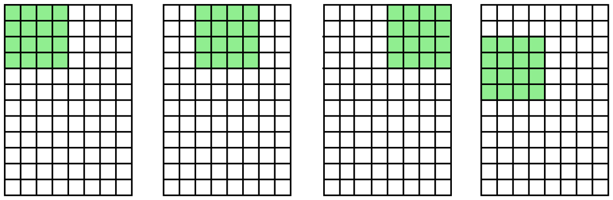
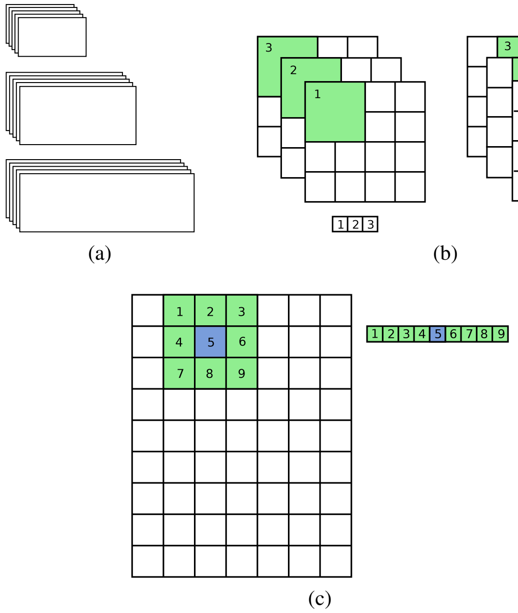


Fig. 2. Four iterations using step of $N/2$ pixels for sliding window. The 3 first steps move to right till the end of the row, then the sliding window is moved $N/2$ pixels underneath at the beginning of the next row.



(a)

(b)

(c)

Fig. 3. Strategies for constructing feature descriptor. (a) frames at multiple resolutions; (b) concatenation of features from consecutive frames; (c) sliding window that concatenates feature descriptors generated around the neighborhood.

(iii) the current block is split into sub-blocks that represent its neighborhood (Figure 3(c)), such that each sub-block is individually processed and, finally, the descriptors are concatenated into a single descriptor.

For training, we considered not to use all frames from the video due to the relative similarity between consecutive frames. The training model was executed using the video frame rate (in Hz) to obtain frames with half second as interval (0.5 Hz). On the other hand, no frame is discarded for predicting the video category.

Two different approaches were initially tested to classify a video as violent: (i) the classification is performed frame-by-frame and the prediction uses 50% as decision threshold; (ii) the decision threshold is calculated based on the value that maximizes the accuracy for training dataset.

C. Data Normalization and Future Reduction

Once the descriptor has been extracted, data scaling normalization is optionally applied to transform the data with zero mean and unit variance. Principal Component Analysis (PCA) technique can also be applied for dimensionality reduction. We evaluate the effectiveness of the normalization and dimensionality reduction of the features in our experiments.

D. Classification

In the training step, we evaluate the use of Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) and linear kernels, as well as decision tree, Adaboost, Gaussian Naïve Bayes and Stochastic Gradient Descent (SGD).

After the training step, each video sequence from the test dataset is individually evaluated. In order to make the final decision, we use a simple strategy for evaluating the prediction for each video frame based on the proportion of violent frames as:

$$\text{output} = \begin{cases} \text{violence}, & \text{if } \mathit{VF} \,/\, (\mathit{VF} + \mathit{NVF}) > T \\ \text{non-violence}, & \text{otherwise} \end{cases} \qquad (2)$$

where *VF* represents the number of violence frames, *NVF* represents the number of non-violence frames, and *T* is a threshold.

To evaluate the accuracy of the classification process, we employ the same protocol as specified by Hassner et al. [1], that is, a k-fold cross-validation protocol. We split the video sequences into k sets, where half the videos in each set portrays violent crowd behavior and half non-violent behavior. In some cases, different videos are originated from the same YouTube clip or the same scene. In such cases, these videos are all included in the same set (the sets were mutually scene exclusive).

In each test, four out five sets are used for training (including SVM , Gaussian Naïve Bayes, Logistic Regression, Random Forest Trees, and PCA or vocabulary generation, when required). Violence labeling is then performed on the remaining set. Results are reported as mean prediction accuracy.

V. Experiments

Our method was implemented using Python programming language and NumPy library. OpenCV was used to preprocess and filter the video sequences, such as Gaussian blur, MoG background subtraction, and HoG descriptor. Classification approaches, such as SVM (using linear and RBF kernels), SGD, Random Forests, and Gaussian Naïve Bayes, were provided by the scikit-learn library.

The classification results were evaluated according to a 5-fold cross-validation protocol, that is, the data set was split into 5 smaller sets, where a model was trained using 4 of the folds as training set. The resulting model was evaluated on the remaining portion of the data. The five resulting outcomes from the folds were then averaged to produce the final estimation.

A. Violent Flows Dataset

Samples extracted from the Violent Flows dataset [1] are illustrated in Figure 4. The dataset contains 246 crowded scenes categorized into two groups, each one with 123 violent and 123 non-violent scenes.

Table I reports our results using CENTRIST-based descriptors, as well as a comparison against other approaches available

Fig. 4. Examples of frames extracted from the Violent Flows dataset [1].

Fig. 5. Examples of frames extracted from the Hockey Fight dataset [2].

in the literature for the Violent Flows dataset. Our experiments employed the following techniques: Histogram Equalization (HEq), Gaussian blur (Blur), MoG background subtraction (MoG), multiple scales (Multiscale), block-[size] for block partitions.

In summary, it is possible to observe that, for some combinations, we achieved superior results compared to the baseline [1] and to other methods available in the literature. For instance, the test using CENTRIST descriptor with PCA and SVM techniques was able to obtain accuracy of 86.16%±2.80%. Even though the concepts involved in the process are relatively simple, the result was higher than other works. On the other hand, the multiscale approach was slightly inferior than its monoscale version, achieving 85.81% ± 2.64% of accuracy, with scales 0.6, 1.0, and 1.4.

The combination between CENTRIST and HOG descriptors produced the best accuracy rate (87.45% ± 2.77%) for the Violent Flows dataset, when applying the background subtraction and PCA for dimensionality reduction, using SVM with RBF kernel for classification. In other experiments without using background subtraction, the results were still representative (86.96% ± 3.86%), such that the use of background subtraction was not sufficiently meaningful. The analysis of several scenes demonstrated that regions of interest can be highly affected by camera motion.

We have also conducted experiments with different classifiers. Using CENTRIST for Violent Flows dataset, the best accuracy rate (87.80% ±1.84%) was obtained using SGD. Using AdaBoost and Random Forests, accuracy rates reached 86.61% and 86.20%, respectively, values quite similar to the result obtained through SVM classifier 85.81%.

Furthermore, we have also evaluated CENTRIST-based descriptors using a sliding window with specific size and step. For a 64×64 window with step 32, we have obtained the results 89.85% and 91.05% for CENTRIST and HOG+CENTRIST, respectively. By applying PCA for reducing the HOG+CENTRIST dimensionality, we achieved 91.46%.

B. Hockey Fights Dataset

Samples extracted from the Hockey Fights dataset [2] are illustrated in Figure 5. The dataset contains 1000 clips categorized into two groups, 500 related to fight scenes and the other half to common hockey scenes. The video sequences were distributed into 5 folds, each one with 100 clips with fight scenes and 100 clips with no fight scenes.

Analogously for the Violent Flows dataset, experiments using the CENTRIST-based descriptors on the HockeyFights dataset achieved promising results, as reported in Table II.

Our approach obtained an accuracy rate of 90.69% applying only PCA and SVM techniques with CENTRIST features over each frame. When background subtraction was previously applied to specific regions of the frame, the resulting accuracy was 91.19% with SVM and with PCA followed by SVM.

Experiments with two other classifiers, Random Forests and AdaBoost, achieved 90.60% and 92.29%, respectively, with HOG+CENTRIST extracted after the application of background subtraction. This latter result represents our best accuracy rate using entire frames for Hockey Fights dataset. On the other hand, Adaboost obtained inferior results when evaluating blocks.

Additionally, when HOG+CENTRIST combination was used with PCA and SVM techniques, the results improved to 90.29%. For our block-based approach, the experiments using CENTRIST and HOG+CENTRIST achieved 91.69% and 92.79%, respectively. This latter value was roughly superior than the result using the entire frames (92.29%).

Although our method has not generated superior results compared to other approaches available in the literature (such as SRC [34] (94.40%), MoWLD [32] (94.20%) and MoIWLD [25] (96.80%)), they are promising.

VI. Conclusions

In this work, we presented a method for detecting violence context from video scenes based on the CENTRIST descriptor. Several experiments were performed on two benchmarks to demonstrate the effectiveness of our method. Although the proposed approach did not overcome the best results reported in the literature, the accuracy rates are very competitive using a approach conceptually simple that is capable of capturing discriminative characteristics for violence classification in video scenes. Furthermore its combination with preprocessing strategies, such as histogram equalization and background subtraction, provided improvements to the descriptor.

Acknowledgments

The authors are thankful to São Paulo Research Foundation (grant FAPESP #2014/12236-1) and Brazilian Council for Scientific and Technological Development (grant CNPq #305169/2015-7) for their financial support.

TABLE I
RESULTS FOR THE VIOLENT FLOWS DATASET [1].

Method	Accuracy (%)
LTP [28]	61.53 ± 0.17
HOG [29]	57.43 ± 0.37
HOF [29]	58.53 ± 0.32
ViF (SVM) [1]	81.30 ± 0.21
SD [30]	85.43 ± 0.21
HOT [31]	82.30
Holistic Features (SVM) [15]	85.53 ± 0.17
ViF + OViF (SVM) [16]	86.00 ± 1.41
ViF + OViF (AdaBoost + SVM) [16]	88.00 ± 2.45
MoSIFT + BoW [2]	57.09 ± 0.37
MoWLD + BoW [32]	82.56 ± 0.19
MoWLD + SparseCoding [32]	86.39 ± 0.15
RVD [24]	82.79 ± 0.19
AMDN [33]	84.72 ± 0.17
MoWLD + KDE + SparseCoding [32]	89.78 ± 0.13
MoIWLD [25]	**93.19** ± **0.12**
CENTRIST (HEq, SVM)	85.75 ± 5.57
CENTRIST (PCA + SVM)	86.16 ± 2.80
CENTRIST (SGD)	87.80 ± 1.84
CENTRIST (MoG, PCA + SVM)	83.73 ± 2.21
CENTRIST (HEq + MoG, SGD)	83.35 ± 1.29
CENTRIST Multiscale (PCA + SVM)	85.81 ± 2.64
CENTRIST Multiscale (SVM)	84.90 ± 5.40
CENTRIST Multiscale (HEq + MoG, SGD)	82.96 ± 2.50
HOG + CENTRIST (SVM)	86.96 ± 3.12
HOG + CENTRIST (AdaBoost)	86.61 ± 4.12
HOG + CENTRIST (Random Forest)	86.20 ± 6.04
HOG + CENTRIST (HEq + MoG, SVM)	87.45 ± 2.77
CENTRIST (HEq, block-96, SVM)	89.88 ± 4.99
CENTRIST (HEq, block-64, SVM)	89.85 ± 3.31
CENTRIST (HEq, block-64, PCA+Adaboost)	87.76 ± 2.43
HOG + CENTRIST (HEq, block-96, SVM)	89.86 ± 2.76
HOG + CENTRIST (HEq, block-72, SVM)	90.26 ± 1.85
HOG + CENTRIST (HEq, block-64, SVM)	91.05 ± 1.64
HOG + CENTRIST (HEq, block-64, PCA+SVM)	**91.46** ± **1.46**

REFERENCES

[1] T. Hassner, Y. Itcher, and O. Kliper-Gross, “Violent flows: Real-time detection of violent crowd behavior,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Providence, RI, USA: IEEE, 2012, pp. 1–6.

[2] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, “Violence detection in video using computer vision techniques,” in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2011, pp. 332–339.

[3] M. F. Alcantara, T. P. Moreira, and H. Pedrini, “Real-time action recognition based on cumulative motion shapes,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 2941–2945.

[4] J. Cai, X. Tang, and G. can Feng, “Learning pose dictionary for human action recognition,” in *International Conference on Pattern Recognition*, vol. 1, Stockholm, Sweden, Aug. 2014, pp. 381–386.

[5] V. Kantorov and I. Laptev, “Efficient feature extraction, encoding and classification for action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014.

[6] B. Liang and L. Zheng, “3D motion trail model based pyramid histograms of oriented gradient for action recognition,” in *International Conference on Pattern Recognition*, vol. 1, Stockholm, Sweden, Aug. 2014, pp. 1952–1957.

[7] J. Wu and J. M. Rehg, “Centrist: A visual descriptor for scene categorization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

TABLE II
Results for the Hockey Fights dataset [2].

Method	Accuracy (%)
HOG + BoW [2]	91.00
HOF + BoW [2]	88.60
MoSIFT + BoW [2]	90.90
MoWLD + BoW [32]	91.90
MoWLD + SparseCoding [32]	93.70 ± 1.68
MoWLD + KDE + SparseCoding [32]	94.20 ± 1.91
MoIWLD + BoW [25]	91.80 ± 1.03
RVD [24]	92.10 ± 1.01
AMDN [33]	89.70 ± 1.13
SRC [34]	94.40 ± 1.07
MoIWLD [25]	**96.80** ± **1.04**
CENTRIST (Blur, PCA + SVM)	90.69 ± 2.55
CENTRIST (Blur + HEq, PCA + SVM)	90.09 ± 2.23
CENTRIST (Blur + HEq + MoG, SVM)	89.89 ± 2.44
CENTRIST (HEq + MoG, PCA + SVM)	91.19 ± 2.33
CENTRIST (Blur + MoG, SVM)	91.19 ± 2.33
HOG + CENTRIST (HEq, SVM)	90.99 ± 2.75
HOG + CENTRIST (HEq, AdaBoost)	92.29 ± 3.68
HOG + CENTRIST (HEq + MoG, AdaBoost)	92.29 ± 3.68
HOG + CENTRIST (HEq + MoG, Random Forest)	90.60 ± 3.02
CENTRIST (HEq, block-96, SVM)	91.69 ± 2.88
HOG + CENTRIST (HEq, block-96, SVM)	**92.79** ± **3.05**

[8] M. F. de Alcantara, T. P. Moreira, and H. Pedrini, “Motion silhouette-based real time action recognition,” in *Iberoamerican Congress on Pattern Recognition*. Havana, Cuba: Springer, 2013, pp. 471–478.

[9] M. Alcantara, T. Moreira, and H. Pedrini, “Real-time action recognition using a multilayer descriptor with variable size,” *Journal of Electronic Imaging*, vol. 25, no. 1, pp. 013 020–013 020, 2016.

[10] M. F. Alcantara, T. P. Moreira, H. Pedrini, and F. Flórez-Revuelta, “Action identification using a descriptor with autonomous fragments in a multilevel prediction scheme,” *Signal, Image and Video Processing*, vol. 11, no. 2, pp. 325–332, 2017.

[11] T. P. Moreira, D. Menotti, and H. Pedrini, “First-person action recognition through visual rhythm texture description,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*. New Orleans, LA, USA: IEEE, 2017, pp. 2627–2631.

[12] S.-H. Cho and H.-B. Kang, “Abnormal behavior detection using hybrid agents in crowded scenes,” *Pattern Recognition Letters*, vol. 44, pp. 64–70, 2014.

[13] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, “Crowded scene analysis: A survey,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[14] J. Shao, K. Kang, C. Change Loy, and X. Wang, “Deeply learned attributes for crowded scene understanding,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4657–4666.

[15] M. Marsden, K. McGuinness, S. Little, and N. E. O’Connor, “Holistic features for real-time crowd behaviour anomaly detection,” in *IEEE International Conference on Image Processing*. Phoenix, AZ, USA: IEEE, 2016, pp. 918–922.

[16] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, “Violence detection using oriented violent flows,” *Image and Vision Computing*, vol. 48, pp. 37–41, 2016.

[17] K. Lloyd, D. Marshall, S. C. Moore, and P. L. Rosin, “Detecting violent crowds using temporal analysis of GLCM texture,” *arXiv preprint 1605.05106*, 2016.

[18] J. Nam, M. Alghoniemy, and A. H. Tewfik, “Audio-visual content-based violent scene characterization,” in *IEEE International Conference on Image Processing*, vol. 1. Chicago, IL, USA: IEEE, 1998, pp. 353–357.

[19] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, “A multimodal approach to violence detection in video sharing sites,” in *20th International Conference on Pattern Recognition*. Istanbul, Turkey: IEEE, 2010, pp. 3244–3247.

[20] I. Laptev and T. Lindeberg, “Space-time interest points,” in *9th International Conference on Computer Vision*. Nice, France: IEEE, 2003, pp. 432–439.

[21] M.-Y. Chen and A. Hauptmann, “MoSIFT: Recognizing Human Actions in Surveillance Videos,” Pittsburgh, PA, USA, Tech. Rep., Sep. 2009.

[22] C. E. Rasmussen, “The infinite Gaussian mixture model,” in *NIPS*, vol. 12, 1999, pp. 554–560.

[23] Z. Zivkovic, “Improved adaptive Gaussian mixture model for background subtraction,” in *17th International Conference on Pattern Recognition,*, vol. 2. IEEE, 2004, pp. 28–31.

[24] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, “A new method for violence detection in surveillance scenes,” *Multimedia Tools and Applications*, vol. 75, no. 12, pp. 7327–7349, 2016.

[25] T. Zhang, W. Jia, X. He, and J. Yang, “Discriminative dictionary learning with motion Weber local descriptor for violence detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 696–709, 2017.

[26] P. KaewTraKulPong and R. Bowden, “An improved adaptive background mixture model for real-time tracking with shadow detection,” in *Video-based Surveillance Systems*. Springer, 2002, pp. 135–144.

[27] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[28] L. Yeffet and L. Wolf, “Local trinary patterns for human action

recognition," in *IEEE 12th International Conference on Computer Vision*. Kyoto, Japan: IEEE, 2009, pp. 492–497.

[29] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA: IEEE, Jun. 2008, pp. 1–8.

[30] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, "Violence detection in crowded scenes using substantial derivative," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*. Colorado Springs, CO, USA: IEEE, 2015, pp. 1–6.

[31] H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *IEEE Winter Conference on Applications of Computer Vision*. Waikoloa Beach, HI, USA: IEEE, 2015, pp. 148–155.

[32] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: a robust motion image descriptor for violence detection," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 1419–1438, 2017.

[33] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," *arXiv preprint 1510.01553*, 2015.

[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.