2018 17th IEEE International Conference on Machine Learning and Applications

Machine Cognition of Violence in Videos using Novel Outlier-Resistant VLAD

Tonmoay Deb
Electrical & Computer Engineering
North South University, Dhaka, Bangladesh
tonmoay.deb@northsouth.edu

Aziz Arman
Electrical & Computer Engineering
North South University, Dhaka, Bangladesh
aziz.arman@northsouth.edu

Adnan Firoze*
Electrical & Computer Engineering
North South University, Dhaka, Bangladesh
adnan.firoze@northsouth.edu

***Abstract*—Understanding highly accurate and real-time violent actions from surveillance videos is a demanding challenge. Our primary contribution of this work is divided into two parts. Firstly, we propose a computationally efficient Bag-of-Words (BoW) pipeline along with improved accuracy of violent videos classification. The novel pipeline's feature extraction stage is implemented with densely sampled Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors rather than Space-Time Interest Point (STIP) based extraction. Secondly, in encoding stage, we propose Outlier-Resistant VLAD (OR-VLAD), a novel higher order statistics-based feature encoding, to improve the original VLAD performance. In classification, efficient Linear Support Vector Machine (LSVM) is employed. The performance of the proposed pipeline is evaluated with three popular violent action datasets. On comparison, our pipeline achieved near perfect classification accuracies over three standard video datasets, outperforming most state-of-the-art approaches and having very low number of vocabulary size compared to previous BoW Models.**

Keywords-Novel Video Encoding, VLAD, Violence Detection, Bag of Words, HOG, Dense Sampling, Violent Flows, OR-VLAD

I. Introduction

Learning human actions from video is an important topic in computer vision. Recently, number of violent activities increased dramatically, from home to street, individual to crowd. Surveillance cameras are being used more than ever [13]. As a result, manual human monitoring on surveillance cameras are becoming obsolete due to inefficiency. Machine intervention on surveillance can eradicate the latency issue considerably. To automate the process, making machines understand violent actions from videos became an important topic. In this paper, we present a novel methodology on this issue, and efficiently improve violent video classification accuracy.

The basic intuition behind comprehending violence is understanding abnormal movement, particularly, abnormal motions in video sequences. Several works have already been done for violent action recognition [9, 15, 21]. Most of them extracted frames from a video clip and analyzed the total video frame by frame. Some of them have already shown satisfactory performance on standard datasets [6, 9]. All the works are classified into two tracks: hand-crafted and based on deep neural networks.

Among several hand-crafted approaches, Histogram of Oriented Gradients (HOG) [27, 28] is frequently used [6]. Other notable approaches are Histogram of Oriented Flow (HOF) [28] and Motion Boundary Histogram (MBH) [25]. The features are robust to camera movement, illumination, low-resolution, lead to highly informative descriptors, and calculated with Spatio-Temporal Interest Point (STIP) [1]. Regions of interest point are represented with HOG, HOF features. Recently, Motion SIFT (MoSIFT) [2] descriptor showed significant performance gain in classifying violent videos [11]. These hand-crafted features are extracted from video frames as local video representations, experimented through Bag-of-Words (BoW) [28] pipeline. One research suggested [10] developing densely sampled descriptors for making representation. Some works [10, 12] focused on 3D HOG, HOF volumes, sampled from sharable subvolumes, and consequently made feature extraction process more efficient compared to STIP [12]. However, according to some experiments [36, 37], BoW model worked better for densely sampled descriptors than key-point based. Some works achieved more efficiency [10, 12] by skipping subsequent frames, because those frames do not represent significant differentiable information. For this research, we extracted dense HOG and HOF descriptors.

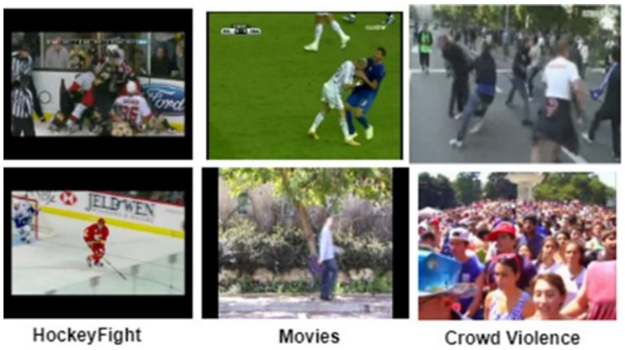

Figure 1. Non-violent (lower) and violent (upper) dataset scenes

In Bag-of-Words (BoW) pipeline, visual vocabulary is generated with k-means clustering from randomly sampled features and contribute feature vector representation with encoding techniques like Fisher Vector (FV) [33] and Vector of Locally Aggregated Descriptors (VLAD) [3]. VLAD has proven to be a computationally efficient feature encoding method [23, 34]. However, though being efficient, standard VLAD descriptor computes only difference between local descriptor and cluster mean. Being in first order statistics, the 'mean' fails representing robust differentiable information of descriptors [35], leading to accuracy lacks. However, some of recent works suggested incorporating higher-order statistics [13, 35] in VLAD. We

978-1-5386-6805-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMLA.2018.00161

addressed this issue and improved VLAD by introducing a novel outlier-resistant encoding by incorporating difference between Median, and Interquartile range on computed visual word vocabulary and local descriptors.

Our core contribution in this paper is that we extract efficient densely sampled features [10, 12] from popular violent action dataset, namely Hockey-Fights, Movies and Crowd Violence [6, 9] illustrated in Figure 1, experiment for an optimum solution on classification. Besides, we propose a novel Outlier-Resistant VLAD (OR-VLAD) encoding for improvement of traditional VLAD in terms of accuracy. Also, we illustrate, how densely extracted, subsampled features encoded by our proposed method achieve higher accuracy with lower computational cost. The experiments were conducted based on densely sampled HOG and HOF features [10, 12]. Further we make our extracted data and encoding codes open to public as supplemental material accompanying this paper[1].

II. Related Works

The first stage of violence detection was limited to several specific information. Nam, Alghoniemy, and Tewfik [4] proposed violent action scene understanding based on blood, flame detection, and sounds of those scenes. Cheng, Chu, and Wu [5] proposed a violent action understanding method based on gunshots, explosions, breaking sounds etc. in videos using Gaussian Mixture Model (GMM) and Hidden Markov Models (HMM). These approaches were largely dependent on audio cues. There is a big chance of misclassification with audio, as crowded scenes mostly carry meaningless information in this context. Moreover, most of the CCTV surveillance cameras do not contain audio. In practical scenarios, high audio dependence raises difficulties.

Another track of classification started along with feature extraction procedures from videos, namely Space-Time Interest Points (STIP) [1] which is an extension of Harris corner detection operator to space-time. These calculated points in spatial and temporal scale, resulting in Histogram of Oriented Gradients (HOG), and Histogram of Optical Flow (HOF) features. Though being robust, Motion SIFT (MoSIFT) [2], an extension to SIFT [8], is computationally expensive. Nievas, Suarez, García, and Sukthankar [6] introduced a new dataset for evaluating violent actions from hockey game, along with a small dataset, where action videos were taken from movies. Both contained a wide variety of scenes and in them half of the clips were violent (fight), and other half clips were non-violent (non-fight). Authors followed standard BoW pipeline, where HOG, HOF [1] and MoSIFT [2] features were extracted and classified using Support Vector Machine (SVM) with faster Histogram Intersection Kernel (HIK) [31]. Hassner, Itcher, and Kliper-Gross [9] introduced Violent Flows dataset, focused on crowd violent scenes from videos. They proposed a novel Violent Flows (ViF) video descriptor, proven to make higher performance in the crowd violence dataset. Later, Xu, Gong, Yang, Wu, and Yao [11] had proposed Sparse Coding over MoSIFT features. However, Gao, Liu, Sun, Wang, and Liu [14] proposed Oriented Violent Flows (OViF) as an extension to existing ViF. Bilinski and Bremond [15] proposed extension of Improved Fisher Vector (IFV) encoding which achieved higher accuracy. Zhang, Jia, He, and Yang [17] proposed Motion Weber Local Descriptor (MoWLD), Souza and Pedrini [18] employed CENTRIST-based features, and Senst, Eiselein, Kuhn, and Sikora [16] introduced feature extraction by using Lagrangian direction. All these followed hand-crafted feature extractions and Bag-of-Words based classification process. Recently, Lu, Yao, Sun, and Zhang [40] proposed a novel locally aggregated descriptors, significantly improved performance of standard histogram-based HOG, HOF and MBH descriptors.

There are several video feature encoding techniques. The most commonly used encoding method is Fisher Vector (FV) [33], and Vector of Locally Aggregated Descriptors (VLAD) [3] for action recognition. As discussed in the literature presented above, FV method showed good encoding performance by soft-assignment. At the same time, it dropped computational efficiency [34]. Recently, Improved Fisher Vectors (IFV) [29] shown to outperform standard FV encoding method by pooling local features from video in global scale. However, Bilinski and Bremond [15] addressed that, it computes spatio-temporal features without knowing their positions, and extended IFV by incorporating spatio-temporal features in the encoding.

Compared to Fisher Vectors, VLAD has proven to be computationally efficient along with good accuracy rate [34]. This work improved VLAD by double assignment approach for deep features [20]. Recently Duta et al. [13] proposed computationally efficient VLAD encoding with shape information (SD-VLAD). This approach had adopted first and second order statistics by calculating standard deviation from vocabulary and local descriptors, followed by calculation of mean Z-Score computed from difference between local descriptors and mean of vocabulary for corresponding visual word. Our work was inspired from this encoding, we focus on computing outlier-resistant OR-VLAD encoding by computing median and interquartile range difference on assigned descriptors in a cluster. This approach outperformed SD-VLAD on experiments with little computational inefficiency. Besides, for classification performance improvement, intra-normalization is applied to VLAD for compressing peak features that can put negative effect on entire dataset as recommended [19].

Recently, Deep Neural Network (DNN) achieved state-of-the-art classification result in image and video recognition with groundbreaking success [7, 20, 21, 22]. Recently, Sudhakaran and Lanz [21] proposed end-to-end Deep Neural Network based model, where Convolutional Neural Network (CNN) was employed to extract deep level features from videos and aggregated the features with a

[1] https://sites.google.com/site/orvladencoding/home

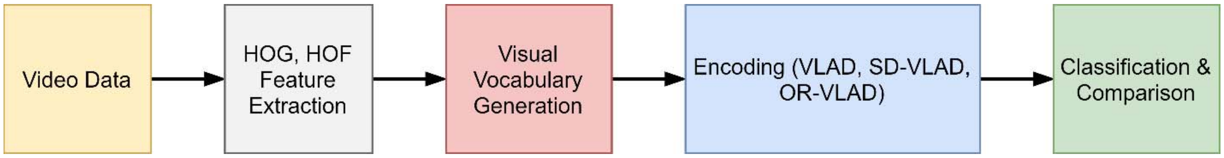
Video Data
HOG, HOF Feature Extraction
Visual Vocabulary Generation
Encoding (VLAD, SD-VLAD, OR-VLAD)
Classification & Comparison

Figure 1. Proposed pipeline of intra-encoding evaluation process

variant of the convolutional gate involved Long Short-Term Memory (LSTM). This outperformed previous approaches. However, our approach, though being hand-crafted, outperformed CNN on Hockey-Fight and Movie datasets classification though being more efficient.

III. Dense Feature Extraction

In this section we present an overview of an efficient, and novel dense feature extraction process as proposed in [10, 12]. For HOG feature extraction, firstly gradient magnitudes of features are computed through HAAR features in vertical and horizontal directions which is proven to do faster computation compared to Gaussian Derivative [27]. Core feature extraction process works on gray video sets, where computed magnitudes generate an output in 2D vector field on each frame, and the vector responses are quantized into orientation bins. The authors had used 8 orientation bins for calculating responses, where each magnitude response fits into 2 neighboring bins according to magnitude value and resulting histogram calculations. Later, the responses are aggregated through spatial and temporal direction. Each block represents size of pixels, and number of frames for a specific location. Each video is spatially divided including number of blocks per orientation aggregation. Then each adjacent block output is concatenated for computing final descriptor output. If block size is 3x3 spatial blocks and 2 temporal blocks with 8 orientations, then the descriptor dimension would be 3x3x2x8=144. Though HOF feature extraction is almost identical to HOG, the core difference is, feature extraction depends on Optical Flow displacement on basis of horizontal direction for one and vertical direction for the other, respectively. There are several ways for computing optical flow, we employed the classic Horn-Schunk [24] as authors suggested. According to the authors, once a block-wise feature is formed, the descriptors of that block are being calculated by doing neighbor-joining. This made the blocks re-usable over orientations. However, 3x3 in spatial domain and 2 in temporal domain make each block being reused for 18 times. Later, responses were aggregated over space with method proposed [38], resulting translation invariant descriptors with interpolation between blocks.

For our experiment in violent video datasets, descriptor extraction block consists of 8x8 pixels and descriptors block consist of 3x3 spatial blocks and two temporal blocks.
For both HOG and HOF, we used 8 orientation bins as per authors' recommendation. Frame Sampling Rate (FSR) was set to 6, resulting number of frames per block to one. Our experiment was performed on Intel Core i3 CPU (2.3GHz), 4.00 GB RAM computer. Table I demonstrate video feature extraction benchmark per frame, clearly represents that our HOG & HOF extraction is faster than other methods.

Table I. Required descriptor extraction time benchmark.

Method	HOG	HOF	ViF [9]	STIP [1]
Per frame time (ms)	7.8	8.3	10	280

IV. Feature Encoding

For representing an input video into single feature vector, standard bag of visual word (BoVW) based classification pipeline was followed. At first, k-means clustering was applied for creating visual vocabulary from set of random 1000 features of each video clip and concatenated to large global representation. For efficiency vs. accuracy tradeoff, classification is evaluated on 50, 100, and 200 visual words.

A. Vector of Locally Aggregated Descriptors

VLAD encoding [3] was proposed by Jégou et al. [34] as non-probabilistic version of Fisher Vector (FV). In pipeline, k number of visual words, created by k-means clustering known as mean centroid vectors of corresponding clusters. Dictionary $D=\mu_1,\mu_{21},\mu_{31},\ldots,\mu_k \in R^{k\times d}$, where each μ represent mean of a visual word, and k is the number of clusters learned by k-means, and d is dimension of the features. From assigned descriptors of encoding, the feature vector is computed through subtracting cluster mean along with summation. $X_i=\{x_1,x_2,x_3,\ldots,x_n\}\in R^{n\times d}$, where n is the number of descriptor x assigned to i -th visual word. VLAD encoding for one cluster is computed as equation 1:

$$VLAD_i=\sum_{j=1}^{n}(x_j-\mu_i) \qquad (1)$$

Encoding of VLAD is obtained by further concentration of the features. For total learned words k, size of encoded vector would be $k\times d$, where d is the dimension of the descriptors. Final representation of VLAD computed with intra-normalization to pick performance as suggested [19].

B. The Outlier-Resistant VLAD Encoding

As per previous section, VLAD compute residuals only based on cluster mean, lacks deeper statistical information. To overcome the shortcomings, several improvements were proposed. Peng, Wang, Qiao, and Peng [35] proposed High-Order Statistics (H-VLAD) for feature encoding. Recently, Duta et al. [13] proposed VLAD encoding with shape information (SD-VLAD). This approach addressed the issue of symmetric, and empty distribution of assigned features to a cluster and introduce standard deviation involvement for improvement. Their robust representation process outperformed VLAD. However, we argue that, taking all assigned descriptor into count for encoding is redundant process because of outlier information existence. Whenever outlier data are used for computing descriptors, encoded result influence negative performance. However, addressing the issue, we proposed a new feature encoding based on median and interquartile range. Median is robust to outlier descriptors, and median- originated interquartile range computes inner outlier-exempted information from the descriptors. Our approach focuses on median of descriptors and global vocabulary assigned to a median cluster. We compute median for every cluster vocabulary followed by the assigned descriptors for each cluster. Later, interquartile range is computed for the descriptors. As per the process, the output result compute highly discriminative and outlier-exempted feature vector. In encoding pipeline, firstly, local descriptor representation is computed to vocabularies with inner dot product between vocabulary and local descriptor [13]. Then we compute mean Z score as per FV and SD-VLAD. Below, equation 2 calculates mean Z-score, computed by residuals divided by standard deviation for a corresponding vocabulary cluster i.

$$V_i^Z = \frac{1}{n}\sum_{j=1}^{n}\frac{(x_j-\mu_i)}{\sigma_i} \tag{2}$$

Here, $Median(K)$ & $IQR(K)$ function calculate median, and interquartile range, accordingly for input vector K as per equation 3 and 4. Due to space shortage, $Median(K)$ is denoted as $M(K)$ in IQR function.

$$Median(K) = \begin{Bmatrix} K\left(\frac{n+1}{2}\right), \; if\ n\ is\ odd \\ \frac{K\left(\frac{n}{2}\right)+K\left(\frac{n+2}{2}\right)}{2}, \; otherwise \end{Bmatrix} \tag{3}$$

$$IQR(K) = \begin{Bmatrix} M\left(K\left(\frac{n+3}{2}, n\right)\right) - M\left(K\left(1, \frac{n-1}{2}\right)\right), \; if\ n\ is\ odd \\ M\left(K\left(\frac{n+2}{2}, n\right)\right) - M\left(K\left(1, \frac{n}{2}\right)\right), \; otherwise \end{Bmatrix} \tag{4}$$

Finally, encoding is computed according to the following equation 5. As per previous section, X_i is the set of features assigned to i-th vocabulary. Features containing all zeros has omitted from X_i to facilitate to avoid 'division by zero'.

$$V_i^M = \frac{Median(X_i) - Median(\mu_i)}{IQR(X_i)} \tag{5}$$

After computing V_i^Z and V_i^M, we concatenate features horizontally, resulting in final descriptor of dimension $2 \times k \times d$. Further, intra-normalization has been applied. In encoding, we employed several statistical information i.e. mean, median, standard deviation, interquartile range for computing robust outlier-resistant encoding. However, in terms of computational cost, as median calculation requires sorted data, time complexity is $O(NlogN)$, where VLAD and SD-VLAD compute with complexity $O(N)$.

V. Classification

Support Vector Machine (SVM) [30] has widely been used for Bag-of-Words classification method as in [6, 11, 14, 38]. For entire classification task, we use Linear SVM (LSVM). Though, most of the related experiments has used Histogram Intersection Kernel (HIK) [31], as per binary classification, we preferred standard linear kernel over HIK. Due to being computationally efficient, LIBLINEAR SVM classifier [32] was employed with custom bias. Further, all classification result was computed with standard 5-fold cross validation.

VI. Experimental Analysis

The experimental pipeline is divided into two stages. In feature extraction stage, efficient dense HOG, and HOF descriptors were extracted. For intra-encoding evaluation, we encoded descriptors with VLAD, SD-VLAD & OR-VLAD on 50, 100 and 200 vocabulary size, and evaluated performance according to accuracy as illustrated pipeline in Figure 2. Additionally, late fusion of extracted HOG, and HOF descriptors was incurred for additional intra-encoding evaluation. Final evaluation of computed results compared with previous state-of-the-art works, based on classification accuracy over three datasets has been presented in Table II.

A. Datasets

We used three standard datasets: Hockey Fights, Movies [6], and Crowd Violence (Violent Flows) [9].

Hockey Fights Dataset [6] is designed for detecting fight in videos. It contains 1000 videos with 500 'fight' and 500 'non-fight' videos of National Hockey League (NHL). All videos are sized 720x576 pixels and contain 50 frames/sec.

Movies Dataset [6] contains 200 clips having wider variety of scenes taken from action movies where 100 are violent, others non-violent, and each are 720x480 pixels.

Crowd Violence Dataset [9] is focused on crowd violent actions. It contains 246 video clips of size 320x240 pixels, where 123 are violent, others non-violent crowd activities. All the videos varied by lengths and contained real-life crowd activity scene sourced from YouTube.

Table II. Classification Result Comparison of Approaches (boldfaced accuracies are highest in all literature till date)

Approach	Hockey Fight	Movies	Crowd Violence
STIP (HOG) + BoW [6]	91.7%	49.0%	57.43±0.37%
STIP (HOF) + BoW [6]	88.6%	59.0%	58.53±0.32%
MoSIFT+HIK [6]	90.9%	89.5%	-
ViF [9]	82.9±0.14%	-	81.3±0.21%
MoSIFT+KDE+Sparse Coding [11]	94.3±1.68%	-	89.05±3.26%
Substantial Derivative [39]	-	96.89± 0.21%	85.43±0.21%
Extended IFV [15]	93.4	99	**96.4**
MoIWLD [17]	96.8±1.04%	-	93.19±0.12%
ViF+OViF [14]	87.5±1.7%	-	88±2.45%
Convolutional + LSTM [21]	97.1±0.55%	**100±0%**	94.57±2.34%
Dense HOG + VLAD	97.6±0.08%	**100±0%**	91.05±2.77%
Dense HOG + SD-VLAD	97.9±0.05%	99.5±0.03 %	90.6±2.80%
Dense HOG + OR-VLAD	**98.2±0.76%**	**100±0%**	93.09±1.14%

B. *Results and Discussion*

For each vocabulary, nine encodings accomplished by HOG, HOF, and HOG+HOF on three encodings, resulting, for a dataset, total 27 training set were computed. All were classified with standard 5-fold cross validation. In most of the cases, HOG with OR-VLAD excelled HOF and HOG+HOF. Table II represent state-of-the-art classification accuracy comparison among the previous works. Upon comparison, OR-VLAD outperformed previous approaches on HockeyFight having 98.2% accuracy, Movies 100%, and competitive 93.09% accuracy on Crowd violence along with superiority compared to other encodings. Figure 3 illustrates Receiver Operating Characteristic (ROC) curve of our encoding-wise comparison along with previous state-of-the-art approaches on the crowd violence dataset. The major reason behind OR-VLAD outperformed other encodings is being outlier-resistant. In most cases crowd-prone violence videos contain highly misleading information about scene, and OR-VLAD handles those as outlier information, which results in significant accuracy improvement.

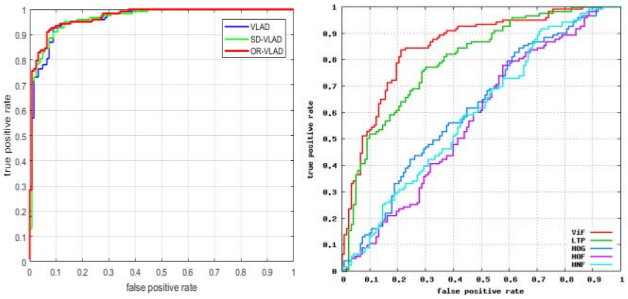

Figure 2. ROC plot. Intra-encoding comparison (**left**), previous accuracies achieved (**right**) on Crowd Violence dataset.

VII. Conclusion

The proposed pipeline implements computationally efficient densely sampled HOG, HOF feature on descriptor extraction stage and we implement a novel new Outlier-Resistant VLAD encoding (OR-VLAD) that outperformed previous approaches in literature. We add that besides detection of violent actions, this encoding can be extended to perform with great efficiency and accuracy in the context of any movement-prone contexts like sports and calamities.

References

[1] I. Laptev, "On Space-Time I nterest Points," Int. J. Comput. Vis., vol. 64, no. 2, pp. 107–123, Sep. 2005.

[2] M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing Human Actions in Surveillance Videos," 2009.

[3] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3304–3311.

[4] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269), 1998, vol. 1, pp. 353–357 vol.1.

[5] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval - MIR '03, Berkeley, California, 2003, p. 109.

[6] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, "Violence Detection in Video Using Computer Vision Techniques," in Computer Analysis of Images and Patterns, 2011, pp. 332–339.

[7] A. Firoze and T. Deb, "Face Recognition Time Reduction Based on Partitioned Faces Without Compromising Accuracy and a Review of State-of-the-art Face Recognition Approaches," in Proceedings of the 2018 International Conference on Image and Graphics Processing, Hong Kong, Hong Kong, 2018, pp. 14–21.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

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

[10] J. R. R. Uijlings, I. C. Duta, N. Rostamzadeh, and N. Sebe, "Realtime Video Classification using Dense HOF/HOG," in Proceedings of International Conference on Multimedia Retrieval, 2014, p. 145.

[11] L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3538–3542.

[12] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, “Video classification with Densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off,” Int J Multimed Info Retr, vol. 4, no. 1, pp. 33–44, Mar. 2015.

[13] I. C. Duta, J. R. R. Uijlings, B. Ionescu, K. Aizawa, A. G. Hauptmann, and N. Sebe, “Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information,” Multimed. Tools Appl., vol. 76, no. 21, pp. 22445–22472, Nov. 2017.

[14] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, “Violence detection using Oriented VIolent Flows,” Image Vis. Comput., vol. 48–49, no. Supplement C, pp. 37–41, Apr. 2016.

[15] P. Bilinski and F. Bremond, “Human violence recognition and detection in surveillance videos,” in 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2016, pp. 30–36.

[16] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, “Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation,” IEEE Trans. Inf. Forensics Secur., vol. 12, no. 12, pp. 2945–2956, Dec. 2017.

[17] T. Zhang, W. Jia, X. He, and J. Yang, “Discriminative Dictionary Learning With Motion Weber Local Descriptor for Violence Detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 3, pp. 696–709, Mar. 2017.

[18] F. D. Souza and H. Pedrini, “Detection of Violent Events in Video Sequences Based on Census Transform Histogram,” in 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2017, pp. 323–329.

[19] R. Arandjelovic and A. Zisserman, “All About VLAD,” in 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1578–1585.

[20] I. C. Duta, T. A. Nguyen, K. Aizawa, B. Ionescu, and N. Sebe, “Boosting VLAD with double assignment using deep features for action recognition in videos,” in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 2210–2215.

[21] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6.

[22] A. Firoze, T. Deb, and R. M. Rahman, “Deep Learning and Data Balancing Approaches in Mining Hospital Surveillance Data,” in Handbook of Research on Emerging Perspectives on Healthcare Information Systems and Informatics, IGI Global, 2018, pp. 140–212.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image Classification with the Fisher Vector: Theory and Practice,” Int. J. Comput. Vis., vol. 105, no. 3, pp. 222–245, Dec. 2013.

[24] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” Artif. Intell., vol. 17, no. 1, pp. 185–203, Aug. 1981.

[25] N. Dalal, B. Triggs, and C. Schmid, “Human Detection Using Oriented Histograms of Flow and Appearance,” in Proceedings of the 9th European Conference on Computer Vision - Volume Part II, Graz, Austria, 2006, pp. 428–441.

[26] H. Wang and C. Schmid, “Action Recognition with Improved Trajectories,” in 2013 IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[27] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, vol. 1, pp. 886–893 vol. 1.

[28] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[29] F. Perronnin, J. Sanchez, and T. Mensink, “Improving the Fisher Kernel for Large-Scale Image Classification,” Plan. Perspect., vol. 143, p. 156, 2010.

[30] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[31] S. Maji, A. C. Berg, and J. Malik, “Classification using intersection kernel support vector machines is efficient,” in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[32] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “LIBLINEAR: A Library for Large Linear Classification,” J. Mach. Learn. Res., vol. 9, no. Aug, pp. 1871–1874, 2008.

[33] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher Kernel for Large-scale Image Classification,” in Proceedings of the 11th European Conference on Computer Vision: Part IV, Heraklion, Crete, Greece, 2010, pp. 143–156.

[34] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating local image descriptors into compact codes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[35] X. Peng, L. Wang, Y. Qiao, and Q. Peng, “Boosting VLAD with Supervised Dictionary Learning and High-Order Statistics,” in Computer Vision – ECCV 2014, 2014, pp. 660–674.

[36] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in BMVC 2009-British Machine Vision Conference, 2009, pp. 124–121.

[37] F. Jurie and B. Triggs, “Creating efficient codebooks for visual recognition,” in Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, 2005, vol. 1, pp. 604–610 Vol. 1.

[38] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha, “Real-Time Visual Concept Classification,” IEEE Trans. Multimedia, vol. 12, no. 7, pp. 665–681, Nov. 2010.

[39] S. Mohammadi, H. Kiani, A. Perina, and V. Murino, “Violence detection in crowded scenes using substantial derivative,” in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1–6.

[40] X. Lu, H. Yao, X. Sun, and Y. Zhang, “Locally aggregated histogram-based descriptors,” *J. VLSI Signal Process. Syst. Signal Image Video Technol.*, vol. 12, no. 2, pp. 323–330, Feb. 2018.