Violence Detection Based on Spatio-Temporal Feature and Fisher Vector

Huangkai Cai[1], He Jiang[1], Xiaolin Huang[1], Jie Yang[1](✉), and Xiangjian He[2]

[1] Institution of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
jieyang@sjtu.edu.cn

[2] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, Australia

Abstract. A novel framework based on local spatio-temporal features and a Bag-of-Words (BoW) model is proposed for violence detection. The framework utilizes Dense Trajectories (DT) and MPEG flow video descriptor (MF) as feature descriptors and employs Fisher Vector (FV) in feature coding. DT and MF algorithms are more descriptive and robust, because they are combinations of various feature descriptors, which describe trajectory shape, appearance, motion and motion boundary, respectively. FV is applied to transform low level features to high level features. FV method preserves much information, because not only the affiliations of descriptors are found in the codebook, but also the first and second order statistics are used to represent videos. Some tricks, that PCA, K-means++ and codebook size, are used to improve the final performance of video classification. In comprehensive consideration of accuracy, speed and application scenarios, the proposed method for violence detection is analysed. Experimental results show that the proposed approach outperforms the state-of-the-art approaches for violence detection in both crowd scenes and non-crowd scenes.

Keywords: Violence detection · Dense Trajectories
MPEG flow video descriptor · Fisher Vector
Linear support vector machine

1 Introduction

Violence detection is to determine whether a scene has an attribute of violence. Violence is artificially defined, and video clips are artificially labelled as 'normal' and 'violence'. Violence detection is considered as not only a branch of action recognition, but also an instance of video classification. Techniques of violence detection can be applied to real life in intelligent monitoring systems and for reviewing videos automatically on the Internet.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11256, pp. 180–190, 2018.
https://doi.org/10.1007/978-3-030-03398-9_16

Early approaches of action recognition are based on trajectories, which need to detect human bodies and track them for video analysis. They are complicated and indirect, because human detection and tracking have to be solved in advance. Recently, the methods based on local spatio-temporal features [16,17] have dominated the field of action recognition. These approaches use local spatio-temporal features to represent global features of videos directly. Moreover, their performance is excellent and robust under various conditions such as background variations, illumination changes and noise. In [11], a Bag-of-Words (BoW) model was used to effectively transform low level features to high level features.

Motivated by the performance of local spatio-temporal features and BoW models, a new framework using Dense Trajectories (DT) [16], MPEG flow video descriptor (MF) [7] and Fisher Vector (FV) [10] for violence detection is proposed as illustrated in Fig. 1. We provide the reasons for why DT and MF are chosen for feature extraction and why FV is chosen for feature coding as follows.

For feature extraction, a variety of feature descriptors based on local spatio-temporal features can be applied. These descriptors include Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) [8], Motion SIFT (MoSIFT) [2], Motion Weber Local Descriptor (MoWLD) [21] and Motion Improved Weber Local Descriptor (MoIWLD) [20]. The applications of these feature descriptors to describe human appearance and motion for violence detection can be found in [11,18,20,21].

For the purpose of extracting more descriptive features to improve the performance of violence detection, DT and MF are utilized for the first time for violence detection in this paper. The interest points that are densely sampled by DT preserve more information than all other features mentioned above. DT is a combination of multiple features including trajectory shape, HOG, HOF and Motion Boundary Histogram (MBH), so it takes the advantages of these features. On the premise of ensuring prediction accuracy, MF improves the computational cost and time consumption compared to DT.

For feature coding, Vector Quantization (VQ) [14] and Sparse Coding (SC) [19] are two commonly used methods for encoding the final representations. VQ votes for a feature only when the feature 'word' is similar to a word in the codebook, so it may result in information loss. SC reconstructs the features by referring to the codebook, preserves the affiliations of descriptors and stores only the zeroth order statistics. The work using SC or its variants for violence detection can be found in [18,20,21].

Compared with VQ and SC, Fisher Vector generates a high dimensional vector that stores not only the zeroth order statistics, but also the first and second order statistics. Moreover, the running time of FV is much less than VQ and SC, hence it is used for feature coding in this paper.

The contributions of this paper are summarized as follows. A novel framework for violence detection is proposed. It uses DT and MF feature descriptors as local spatio-temporal features and utilizes FV for feature coding. Some tricks, that PCA, K-means++ and codebook size, are applied to improve the performance of violence detection. Our proposed framework of violence detection is analysed from various aspects including accuracy, speed and application scenarios. Experimental results demonstrate that the proposed approach outperforms the state-of-the-art techniques on both crowd and non-crowd datasets in terms of accuracies.

The rest of this paper is organized as follows. In Sect. 2, we will elaborate the proposed framework including Dense Trajectories, MPEG flow video descriptor and Fisher Vector. In Sect. 3, the experimental results in crowd scenes and non-crowd scenes will be showed and analysed. In Sect. 4, conclusions will be discussed.

2 Methodology

This article proposes a novel framework of violence detection using Dense Trajectories (DT), MPEG flow video descriptor (MF) and Fisher Vector (FV) as illustrated in Fig. 1. Firstly, from the violent video clips for training and testing, DT or MF feature vectors are extracted and they describe trajectory shape, appearance, motion and motion boundaries. Secondly, PCA is applied to eliminate redundant information after low level representations are generated. Thirdly, testing videos are encoded as high level representations by FV according to the codebook generated by Gaussian Mixture Models (GMM). Finally, linear SVM is employed to classify the videos into two categories of normal patterns and violence patterns. The algorithm for violence detection in videos based on this framework is detailed in the following subsections.

2.1 Dense Trajectories and MPEG Flow Video Descriptor

Dense Trajectories proposed in [16] is an excellent algorithm of feature extraction for action recognition. DT extracts four types of features that are trajectory shape, HOG, HOF and MBH. These features are combined to represent a local region in the visual aspects of trajectory shape, appearance, motion and motion boundaries.

MPEG flow video descriptor proposed in [7] is an efficient video descriptor which uses motion information in video compression. The computational cost of MF is much less than DT, because the spare MPEG flow is applied to replace the dense optical flow. Furthermore, there exists only minor reduction in the performance of video classification in contrast to DT. The design of MPEG flow video descriptor follows Dense Trajectories except features based on trajectory shape.

The feature descriptor of DT is a 426 dimensional feature vector, which contains a 30 dimensional trajectory shape descriptor, a 96 dimensional HOG descriptor, a 108 dimensional HOF descriptor and a 192 dimensional MBH descriptor. Compared to DT descriptor, MF is a 396 dimensional feature vector without a 30 dimensional trajectory shape descriptor. As types of feature descriptor, DT and MF are pretty descriptive and robust because of the combination of multiple descriptors.

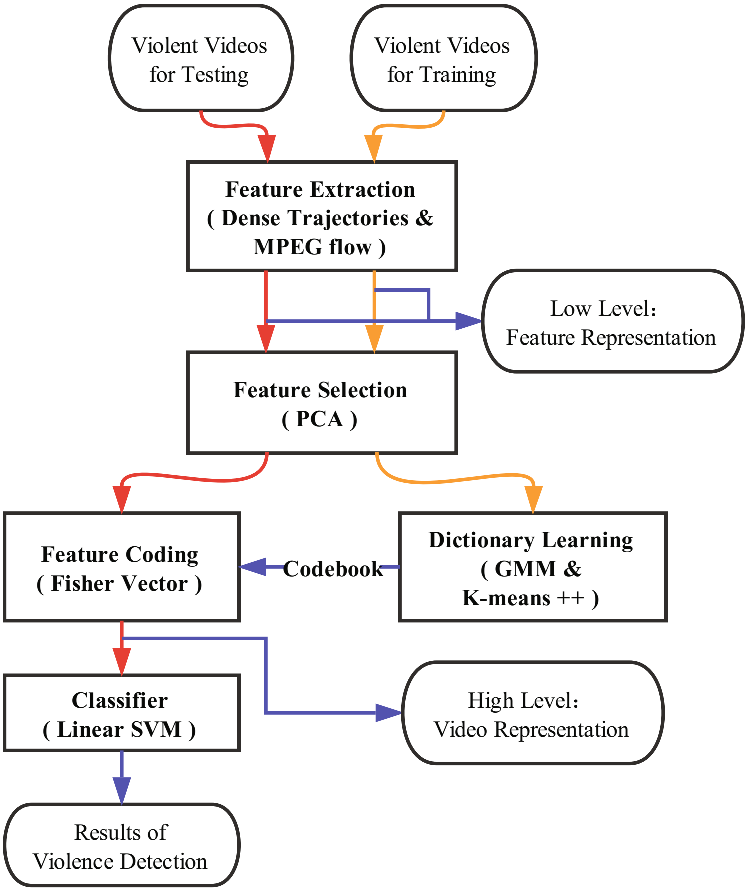
Violent Videos for Testing
Violent Videos for Training
Feature Extraction (Dense Trajectories & MPEG flow)
Low Level: Feature Representation
Feature Selection (PCA)
Feature Coding (Fisher Vector)
Codebook
Dictionary Learning (GMM & K-means ++)
Classifier (Linear SVM)
High Level: Video Representation
Results of Violence Detection

Fig. 1. The proposed framework of violence detection

2.2 Principal Component Analysis

Principal Component Analysis [9,15] is a statistical algorithm for dimensionality reduction. Due to the high dimension of DT (426 dimensional) and MF (396 dimensional), PCA is utilized to reduce the dimension of feature vectors in order to speed up the process of dictionary learning and improve the accuracy of classification. In addition, a whitening process usually follows the PCA, which ensures all features to have the same variance. The transform equation is illustrated as follows.

$$x_{PCA} = \Lambda U^T x_{Original} \tag{1}$$

where $x_{Original} \in R^M$ denotes an original feature, $x_{PCA} \in R^N$ denotes the PCA-Whiten result, $U \in R^{M \times N}$ is the transform matrix of the PCA algorithm, $\Lambda \in R^{N \times N}$ is the whitening diagonal matrix.

2.3 Fisher Vector

Fisher Vector [12,13] is an efficient algorithm for feature coding. It is derived from a fisher kernel [6]. Moreover, FV is usually employed to encode a high level representation of a high dimension for image classification [10]. Both of the first and second order statistics are encoded leading to a high separability of the final feature representations. The FV algorithm is described as follows.

GMM is employed to learn the codebook, which uses generative models to describe the probability distribution of feature vectors. Let $X = \{x_1, \, ... \, , x_N\}$ be a set of D dimensional feature vectors processed through the DT and PCA algorithms, where N is the number of feature vectors. The density $p(x|\lambda)$ and the k-th Gaussian distribution $p_k(x|\mu_k, \varSigma_k)$ are defined as:

$$p(x|\lambda) = \sum_{k=1}^{K} \omega_k p_k(x|\mu_k, \varSigma_k), \tag{2}$$

and

$$p_k(x|\mu_k, \varSigma_k) = \frac{exp[-\frac{1}{2}(x-\mu_k)^T \varSigma_k^{-1}(x-\mu_k)]}{(2\pi)^{D/2}|\varSigma_k|^{1/2}}, \tag{3}$$

where K denotes the mixture number, $\lambda = (\omega_k, \; \mu_k, \; \varSigma_k \, : \, k = 1, \, ... \, , K)$ are the GMM parameters that fit the distribution of the feature vectors, ω_k denotes the mixture weight, μ_k denotes the mean vector and $\varSigma_k$ denotes the covariance matrix.

The optimal parameters forming λ of GMM are learned by the Expectation Maximization (EM) algorithm [3]. Furthermore, the initial values of these parameters have an important influence on the final codebook, so k-means++ [1] results are calculated as the initial values.

In the following equation, y_{ik} represents the occupancy probability, which is the soft assignment of the feature descriptor x_i to Gaussian k:

$$y_{ik} = \frac{exp[-\frac{1}{2}(x_i-\mu_k)^T \varSigma_k^{-1}(x_i-\mu_k)]}{\sum\limits_{t=1}^{K} exp[-\frac{1}{2}(x_i-\mu_t)^T \varSigma_k^{-1}(x_i-\mu_t)]}. \tag{4}$$

Then, the gradient vector $g_{\mu,d,k}^X$ with respect to the mean μ_{dk} of Gaussian k and the gradient vector $g_{\sigma,d,k}^X$ with respect to the standard deviation σ_{dk} of Gaussian k could be calculated. Their mathematical expressions are:

$$g_{\mu,d,k}^X = \frac{1}{N\sqrt{\omega_k}} \sum_{i=1}^{N} y_{ik} \frac{x_{di}-\mu_{dk}}{\sigma_{dk}}, \tag{5}$$

and

$$g_{\sigma,d,k}^X = \frac{1}{N\sqrt{2\omega_k}} \sum_{i=1}^{N} y_{ik} [(\frac{x_{di}-\mu_{dk}}{\sigma_{dk}})^2 - 1], \tag{6}$$

where $d = 1, \ldots, D$ for D representing the dimension of the feature vectors.

Finally, the Fisher Vector is the concatenation of $g^X_{\mu,d,k}$ and $g^X_{\sigma,d,k}$ for $k = 1, \ldots K$ and $d = 1, \ldots, D$, and it is represented by

$$\Phi(X) = [g^X_{\mu,d,k},\ g^X_{\sigma,d,k}]. \tag{7}$$

Therefore, the final representation of a video is $2 \times K \times D$ dimensional.

2.4 Linear Support Vector Machine

Before applying the video representations in the linear SVM, the power and ℓ_2 normalization are applied to the Fisher Vector $\Phi(X)$ as shown in [13]. Then, the linear SVM [4] is used for the violence classification of each video encoded by FV.

3 Experiments

3.1 Datasets

In our experiments, two public datasets are applied to detect whether a scene has a characteristic of violence. These datasets are Hockey Fight dataset (HF dataset) [11] and Crowd Violence dataset (CV dataset) [5]. HF dataset shows non-crowd scenes, while CV dataset shows crowd scenes. The validity of the proposed framework for violence detection will be verified in both crowd scenes and non-crowd scenes. Some frame samples taken from them are displayed in Fig. 2. The datasets are introduced briefly below.

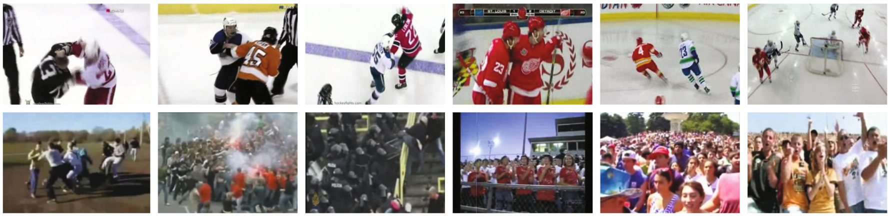

Fig. 2. Frame samples from the Hockey Fight dataset (first row) and the Crowd Violence dataset (second row). The first row shows non-crowd scenes, while the second row shows crowd scenes. The left three columns show violent scenes, while the right three columns show non-violent scenes.

Hockey Fight Dataset. This dataset contains 1000 video clips from ice hockey games of the National Hockey League (NHL). There are 500 video clips labelled as violence, while other 500 video clips are manually labelled as non-violence. The resolution of each video clip is 360 × 288 pixels.

Crowd Violence Dataset. This dataset contains 246 video clips of crowd behaviours, and these clips are collected from YouTube. It consists of 123 violent clips and 123 non-violent clips with a resolution of 320 × 240 pixels.

Table 1. Violence detection results using Sparse Coding (SC) on Hockey Fight dataset

Visual words	MoSIFT + SC [18]		MoWLD + SC [21]	
	ACC	AUC	ACC	AUC
50 words	85.4	0.9211	89.1	0.9318
100 words	88.4	0.9345	90.5	0.9492
150 words	89.6	0.9407	**92.4**	0.9618
200 words	89.6	0.9469	93.1	0.9708
300 words	91.8	0.9575	93.5	0.9638
500 words	92.3	0.9655	93.3	0.9706
1000 words	93.0	0.9669	93.7	0.9781
Visual words	DT + SC		MF + SC	
	ACC	AUC	ACC	AUC
50 words	90.3	0.9542	**91.4**	**0.9564**
100 words	91.6	0.9662	**92.7**	**0.9700**
150 words	91.2	0.9621	92.1	**0.9744**
200 words	92.3	0.9718	**93.5**	**0.9766**
300 words	92.5	0.9759	**93.9**	**0.9792**
500 words	92.4	0.9776	**94.4**	**0.9823**
1000 words	94.4	0.9831	**94.9**	**0.9868**

3.2 Experimental Settings

In feature extraction, experiments are conducted based on three feature descriptors, which are MoSIFT [2] (256 dimensional), Dense Trajectories (DT) [16] (426 dimensional) and MPEG flow video descriptor (MF) [7] (396 dimensional).

For feature selection, PCA is utilized to reduce the abovementioned three types of features to the same dimension of $D = 200$.

For dictionary learning, $100,000$ features are randomly sampled from the training set. For GMM training, k-means++ [1] is used to initialize the covariance matrix of each mixture. It is an important trick for improving the final performance and making the results more stable. The mixture number of GMMs is set to be $K = 256$.

After the codebook is generated, the results using FV are compared with the results using SC in feature coding. The parameter settings of SC are according to those in [18]. The final feature vectors of videos are powered and ℓ_2-normalized.

Finally, the linear SVM [4] is employed for classification of the testing videos, and the penalty parameter is set to be $C = 100$.

5-fold cross validation is used for evaluating the accuracies of video classification. The experimental results are reported in terms of mean prediction accuracy (ACC) and the area under the ROC curve (AUC).

3.3 Experimental Results on Hockey Fight Dataset

We perform a series of experiments for testing the superiority of 4 types of feature descriptors. The 4 types of features are MoSIFT, MoWLD [21], DT and MF, and they are used together with SC on the Hockey Fight dataset. The results from DT + SC and MF + SC are compared with those using the methods recently developed in [18,21]. Furthermore, in order to assess the effect of the codebook size, we set 7 groups of experiments using SC, where the codebook sizes range from 50 words to 1000 words.

Table 2. Violence detection results using Fisher Vector (FV) on Hockey Fight dataset

Methods	ACC	AUC
MoSIFT + FV	93.8	0.9843
DT + FV	94.7	0.9830
MF + FV	95.8	0.9897
MoSIFT + PCA + FV	93.6	0.9859
DT + PCA + FV	95.2	0.9849
MF + PCA + FV	**95.8**	**0.9899**

As shown in Table 1, it is firmly convinced that the features of DT and MF are more effective and discriminative in contrast with the MoSIFT and MoWLD features. DT and MF features are introduced to violence detection for the first time, but they show strong adaptability to non-crowd scenes. In overall consideration of ACC and AUC values, the performance of MF features is the best in these experiments.

The experimental results also indicate that the performance of these algorithms improves with the increase of visual words, i.e., the codebook size contributes to the accuracy of violence detection. In practical application, time consumption will increase if the codebook size expands. So, we can utilize codebook size as a trick to trade off prediction accuracy and time consumption.

FV is applied as an algorithm for feature coding on the Hockey Fight dataset. The performance of FV demonstrated in Table 2 is superior to the performance of SC shown in Table 1. Furthermore, the employment of PCA contributes to the improvement of ACC and AUC, as particularly seen in the results using DT.

In summary, our proposed framework of violence detection, MF + PCA + FV, outperforms the state-of-the-art methods in non-crowd scenes.

3.4 Experimental Results on Crowd Violence Dataset

We compare our proposed algorithm with various state-of-the-art methods including ViF [5], MoSIFT + SC [18], MoWLD + SC [21] and MoIWLD +

Table 3. Violence detection results of various methods on Crowd Violence dataset

Methods	ACC	AUC
ViF [5]	81.30	0.8500
MoSIFT + SC [18]	80.47	0.9008
MoWLD + SC [21]	86.39	0.9018
MoIWLD + SRC [20]	93.19	0.9508
MF + SC	90.63	0.9630
DT + SC	91.45	0.9664
MF + FV	89.83	0.9672
DT + FV	93.50	**0.9889**
MF + PCA + FV	91.89	0.9789
DT + PCA + FV	**95.11**	0.9866

SRC [20] on the Crowd Violence dataset. The codebook size of the compared methods is set to be 500 visual words.

Obviously, our FV based method outperforms the state-of-the-art approaches as shown in Table 3. Moreover, the utilization of PCA effectively improves the accuracy of violence detection.

In crowd scenes, the performance of MF features is inferior to DT features. Because, the information which MF preserves is insufficient due to video compression.

3.5 Analysis of Violence Detection

Comparative analysis of accuracy and speed for violence detection is as shown in Table 4. Speed means that how many frame pictures can be processed per second by different algorithms of feature extraction. We mainly analyse our proposed framework that DT + PCA + FV and MF + PCA + FV in different scenes.

Table 4. Comparative analysis of accuracy and speed for violence detection

Methods	HF dataset		CV dataset		Speed (fps)
	ACC	AUC	ACC	AUC	
DT	95.20	0.9849	95.11	0.9866	1.2
MF	95.80	0.9899	91.89	0.9789	168.4

If time consumption becomes a primary consideration, the framework based on MF will be the optimal choice in both crowd scenes and non-crowd scenes.

Nevertheless, the diversity of application scenarios will result in different options if prediction accuracy is major concerned. The prediction accuracy of

MF is superior to DT in non-crowd scenes, while DT outperforms MF in crowd scenes.

4 Conclusion

This paper has proposed a novel framework of violence detection using Dense Trajectories, MPEG flow video descriptor and Fisher Vector. Firstly, the experimental results have shown that DT and MF as types of discriminative feature descriptors outperform other commonly used features for violence detection. Secondly, FV as an excellent feature coding algorithm has been proven to be superior to Sparse Coding. Thirdly, some tricks including PCA, K-means++ and codebook size have contributed to the improvement of accuracy and AUC values in violence detection. Fourthly, our proposed framework of violence detection was analysed in overall consideration of accuracy, speed and application scenarios. Fifthly, the performance of the proposed method was better than the state-of-the-art techniques for violence detection in both crowd scenes and non-crowd scenes. As our future work, whether DT, MF and FV are suitable for other tasks of video analysis will be further researched.

Acknowledgements. This research is partly supported by NSFC, China (No: 61572315, 6151101179) and 973 Plan, China (No. 2015CB856004).

References

1. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Eighteenth ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
2. Chen, M.Y., Hauptmann, A.: MoSIFT: recognizing human actions in surveillance videos. Ann. Pharmacother. **39**(1), 150–152 (2009)
3. Dempster, A.P.: Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
4. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. J. Mach. Learn. Res. **9**(9), 1871–1874 (2008)
5. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6 (2012)
6. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: International Conference on Neural Information Processing Systems, pp. 487–493 (1998)
7. Kantorov, V., Laptev, I.: Efficient feature extraction, encoding, and classification for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2593–2600 (2014)
8. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
9. Martinsson, P.G., Rokhlin, V., Tygert, M.: A randomized algorithm for the decomposition of matrices. Appl. Comput. Harmon. Anal. **30**(1), 47–68 (2011)

10. Nchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
11. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_39
12. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
13. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
14. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, p. 1470 (2003)
15. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. J. Neural Comput. **11**(2), 443–482 (1999)
16. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
17. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision, pp. 3551–3558 (2014)
18. Xu, L., Gong, C., Yang, J., Wu, Q., Yao, L.: Violent video detection based on MoSIFT feature and sparse coding. In: IEEE Conference on Acoustics, Speech and Signal Processing, pp. 3538–3542 (2014)
19. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)
20. Zhang, T., Jia, W., He, X., Yang, J.: Discriminative dictionary learning with motion weber local descriptor for violence detection. IEEE Trans. Circuits Syst. Video Technol. **27**(3), 696–709 (2017)
21. Zhang, T., Jia, W., Yang, B., Yang, J., He, X., Zheng, Z.: MoWLD: a robust motion image descriptor for violence detection. Multimed. Tools Appl. **76**(1), 1–20 (2017)