Violence Video Detection by Discriminative Slow Feature Analysis

Kaiye Wang[1], Zhang Zhang[2], and Liang Wang[2]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Nowadays, Internet makes it easy for us to share all kinds of information. However, violent content in web has harmful influence on those who lack proper judgment, especially teenagers. This paper presents an approach for detecting violence in videos, where Discriminative Slow Feature Analysis (D-SFA) is introduced to learn slow feature functions from dense trajectories derived from videos. Afterwards, with the learnt slow feature functions, the Accumulated Squared Derivative (ASD) features are extracted to represent videos. Finally, a linear support vector machine (SVM) is trained for classification. We also construct a Violence Video (VV) dataset which includes 200 violence samples and 200 non-violence samples collected from Internet and movies. The experimental results on the newly established dataset demonstrate the effectiveness of the proposed method.

Keywords: violence detection, discriminative slow feature analysis, dense trajectories.

1 Introduction

With the rapid growth of social network websites, such as Facebook, Twitter, and Youtube, etc., a large number of videos are being uploaded everyday. As we enjoy useful information conveniently from these websites, some videos including violence can also be accessed by users. For those who lack proper judgment, e.g., children, teenagers, exposure to violent content can lead to aggressive behavior or even crime resulting from imitating what they see in those harmful sources. It is therefore obvious that the need of protection of such sensitive social groups, using efficient, automatic, content-based violence detectors, is imperative.

Violence detection is very important, although it's not a hot topic in computer vision. Some approaches have already been proposed to address this problem. In [1], Giannakopoulos et al. used eight audio features, both from the time and frequency domain, as the input to a binary classifier which decides the video content with respect to violence. Then, they extended their work to a multi-class classification problem using Bayesian networks [2]. Gong et al. [3] proposed a three-stage method integrating visual and auditory cues to detect violent scenes

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 137–144, 2012.

in movies. Their approach detects various high-level audio effects related to violence to improve the accuracy. Lin and Wang [4] presented a weakly-supervised audio violence classifier combined with a motion, explosion and blood video classifier in a co-training way to detect violent scenes in movies. Giannakopoulos et al. [5] put forward a method for violence detection in movies based on the statistics of audio features, average motion and motion orientation variance features in video.

In summary, these previous works generally require audio cues for detecting violence. However, audio is not always available in real applications, particularly in some visual surveillance scenarios. Moreover, audio sometimes provides inconsistent information so that the detection result is not reliable. Thus, in this paper, we focus on detection of violence in videos using single visual cues in those cases.

In this paper, we present a new approach for detecting violence in the Slow Feature Analysis (SFA) [6] framework. The discriminative SFA (D-SFA) learning [7] which takes the supervised information into original SFA is adopted for violence detection. The main contribution of this work is two-fold: one is the establishment of a new violence video dataset, including 200 positive and 200 negative samples, which provides a public platform for evaluating the violence detection algorithms, and the other is the D-SFA based violence detection algorithm, where dense trajectories are extracted to form the input of D-SFA learning. Instead of the representation based on local spatial-temporal interest points, dense trajectories obtained by tracking interest points through image sequences can provide richer motion information of the video [8]. Experiments on the newly built database demonstrate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. In Section 2, we present our approach for violence video detection in details. In Section 3, we introduce the newly built violence video database, and the experimental results. Section 4 concludes this paper.

2 Violence Video Detection

The flowchart of our method is presented in Fig.1. Firstly, dense trajectories are extracted by computing dense optical flow field. Then, the D-SFA is performed to obtain discriminative slow feature functions from training cuboids sampled from dense trajectories. Afterward, the ASD feature is calculated to represent each video clip. Finally, we use a linear SVM for classification.

2.1 Extraction of Dense Trajectories

Feature trajectories have shown to be efficient for representing videos. To extract dense trajectories, we use the method in [8] which is robust to fast irregular motions and shot boundaries. We sample dense points from each frame and track them based on displacement information from a dense optical flow field. Dense trajectories are extracted for multiple spatial scales. Feature points are

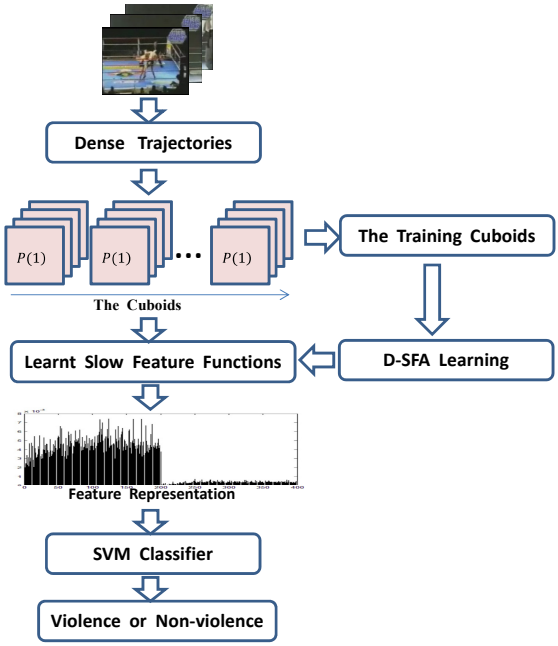


Fig. 1. The overview of our approach

sampled on a grid spaced by 5 pixels and tracked in each scale separately. We used 8 spatial scales spaced by a factor of $1/\sqrt{2}$. Each point $P_t = (x_t, y_t)$ at frame t is tracked to the next frame $t+1$ by median filtering in a dense optical flow field $\omega = (u_t, v_t)$,

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \tag{1}$$

where M is the median filtering kernel, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t). Once the dense optical flow field is computed, points can be tracked very densely without additional cost. Points of subsequent frames are concatenated to form a trajectory:$(P_t, P_{t+1}, P_{t+2}, ...)$. To extract dense optical flow, we use the algorithm of Farneback [9] as implemented in the OpenCV library.

2.2 Collection of Training Cuboids

Given a trajectory of length l, in each frame, a local patch centered at the corresponding trajectory point with the size of $h \times w$ is obtained. Then, a cuboid (or patch sequence) is constructed with the size of $h \times w \times l$ ($15 \times 15 \times l$ in this paper). According to [7], we concatenate $\Delta t = 3$ successive patches to form an input vector of SFA learning, then each cuboid is represented by a vector sequence with the time length of $l - \Delta t + 1$. Fig.2 presents the construction process of cuboids.

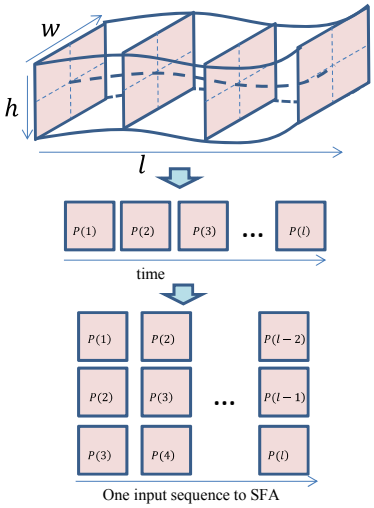


Fig. 2. The construction process of cuboids

2.3 Slow Feature Function Learning

Slow feature analysis (SFA) [6] extracts slowly varying features from a quickly varying input signal. Recently, Zhang and Tao [7] take into account the label information of training samples and propose D-SFA for action recognition. Mathematically, D-SFA is defined as follows:

Given C classes of I-dims input signals $\{\mathbf{x}_c(t) = [x_{c1}(t), ..., x_{cI}(t)] | c \in \{1, ...C\}\}$ with $t \in [t_0, t_1]$ indicating time, for the c-th class, D-SFA finds a set of J-dims input-output functions $\mathbf{g}_c(\mathbf{x}) = [g_{c1}(\mathbf{x}), ..., g_{cJ}(\mathbf{x})]^T$ to minimize

$$\Delta(g_{cj}(\mathbf{x}_c)) - \gamma * \Delta(g_{cj}(\mathbf{x}_{c'})) \tag{2}$$

subject to:

$$\langle g_{cj}(\mathbf{x}_{c\cup c'})\rangle_t = 0 \quad (i.e., zero\ \ mean) \tag{3}$$

$$\left\langle [g_{cj}(\mathbf{x}_{c\cup c'})]^2 \right\rangle_t = 1 \quad (i.e., unit\ \ variance) \tag{4}$$

$$\forall j' < j : \langle g_{cj'}(\mathbf{x}_{c\cup c'}) g_{cj}(\mathbf{x}_{c\cup c'})\rangle_t = 0 \quad (i.e., decorrelation) \tag{5}$$

where $\langle y\rangle_t$ is the mean of signal y over time and γ is the tradeoff parameter.

If the transformation is linear, i.e., $g_{cj}(\mathbf{x}) = w_{cj}^T \mathbf{x}$, wherein $\mathbf{x}$ is the input and w_{cj} is the weight, the solution of D-SFA is equivalent to the generalized eigenvalue problem [6]. To obtain nonlinear slow feature functions, we further perform the nonlinear expansion according to [7] before the D-SFA learning.

2.4 Feature Representation

According to Section 2.2, for each trajectory with the length of l, one cuboid with the time length of $l-\Delta t+1$ is obtained by reformatting the local patch sequence. With the learnt slow feature functions, each input sequence is transformed to a new vector sequence with the size of $K \times (l - \Delta t + 1)$, wherein K is the number of the learnt slow feature functions. For cuboid C_i and slow feature function F_j, the squared derivative $v_{i,j}$ is

$$v_{i,j} = \frac{1}{l - \Delta t} \sum_{t=1}^{l-\Delta t} [C_i(t+1) \otimes F_j - C_i(t) \otimes F_j]^2 \tag{6}$$

where $\otimes$ is the transformation operation (if linear transformation, the operation is inner product; otherwise, before the inner product, C_i is firstly transformed by a non-linear expansion). The ASD feature is formed by accumulating the squared derivatives over all cuboids,

$$f_{ASD} = \sum_{i=1}^{N} V_i \tag{7}$$

where N is the total number of cuboids in current snippet, and $V_i =< v_{i,1}, v_{i,2}, ..., v_{i,K} >^T$.

The number of cuboids collected in a snippet may differ from that in another snippet, so we perform the $L1$ normalization to normalize the feature vector. After the computation of the ASD feature, the LIBSVM [10] with the linear kernel is adopted for violence/non-violence classification.

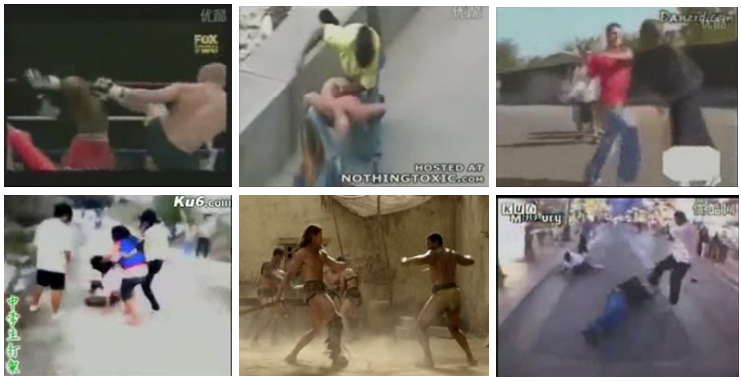
3 Experimental Results

3.1 Evaluation Dataset

To our knowledge, there is not a well-known public database for violence detection. To evaluate the performance of our approach, we collect 200 violence videos and 200 non-violence videos from movies and the Internet to form a dataset named the Violence Video (VV) dataset. The videos cover a wide range of categories, e.g., news, surveillance scenes, sports, movie scenes, etc. The duration of each video is generally less than 60s and the resolution has been normalized to 320×240 pixels. Key frames representing samples of violence and non-violence videos are shown in Fig. 3, respectively. For the research purpose, the VV dataset will be published online soon.

3.2 Experimental Setting and Results

For D-SFA learning, 5000 training trajectories per class are randomly selected and the tradeoff parameter γ is set as 0.2. In this paper, we use the quadratic expansion to learnt non-linear slow feature functions. The first 200 functions per



(a) violence video samples

(b) non-violence video samples

Fig. 3. Video samples in the VV dataset

class are obtained for computing ASD features. Accordingly, the dimensionality of the ASD feature is equal to 400 (i.e., 200×2 for two-class problem here). Here, the ASD feature is computed by three strategies:

(a) For each video, one ASD feature is calculated over all trajectories in the video.

(b) For each frame in one video, one ASD feature is calculated over all trajectories ending in the frame.

(c) For each video, ASD features is calculated over all trajectories ending in a sliding window ($window\ size = 10$ and $step\ length = 5$).

For the last two feature computing strategies (b) and (c) by which one class label can be obtained per frame or every 10 frames, the majority voting rule is accordingly adopted to determine the label of a whole video (i.e., violence or non-violence).

For SVM classification, each time the dataset is splited into a test set with 20 violence videos and 20 non-violence videos selected randomly, and a training set with the remaining 360 videos. We calculate the average performance over

Table 1. Performance of the D-SFA+ASD feature for violence detection

Methods	Accuracy (%)
D-SFA+ASD+SVM (a)	91.92 ± 4.29
D-SFA+ASD+SVM+Voting (b)	93.07 ± 4.64
D-SFA+ASD+SVM+Voting (c)	93.60 ± 3.91

100 random splits. The mean classification accuracies and the corresponding standard deviations are shown in Table 1. From this table, the strategy (c) achieves the highest accuracy, which suggest that the dense trajectories in a small sliding window (e.g., 10 frames) are sufficient to calculate a stable statistical feature (ASD feature), while the voting can further enhance the robustness of the final decision.

3.3 Comparison and Analysis

For comparison, we perform Spatial Temporal Interest points (STIP) + HOG/HOF features [11], which is a state-of-the-art method for action recognition. Firstly, a set of STIPs are detected by 3D Harris corner detector. Then a codebook (400 codes) is learned by the k-means clustering from 200,000 sampled STIPs. Finally, the Bag-of-Words (BoW) model is calculated for representing each video. The results of the BoW method with 100 splits of training/test sets is shown in Table 2.

Table 2. Comparison

Methods	Performance (%)
STIP+HoG/HoF+BoW+SVM [11]	69.08 ± 6.38
D-SFA+ASD+SVM (a)	91.92 ± 4.29

As shown in Table 2, the proposed method achieves much higher accuracies than the BoW method. The BoW method fails to obtain a high performance, because the intra-class variance in the VV dataset is very large, where both motion directions and magnitudes in violence videos are varying irregularly. In contrast, D-SFA extracts discriminative slow variant motion patterns hidden in dense trajectories and the ASD features can effectively encode the magnitude of the motion patterns existing in violence videos. Therefore the D-SFA+ASD can achieve better performance for violence detection.

4 Conclusions

In this paper, we have proposed a method for detecting violence by slow feature analysis. The D-SFA is performed to learn the slow feature functions that can

encode the discriminative information of violence and non-violence videos. The ASD features are computed to represent the video and a linear SVM is used for classification. To evaluate the effectiveness of our proposed method, the experiments have been conducted on a newly built violence video dataset. The results have shown that our approach achieves a promising performance.

In future work, we will continue to refine and enlarge our violence video database by collecting more diverse and representative violence videos, so that we can evaluate the violence detector more fairly.

Acknowledgments. This work is jointly supported by National Natural Science Foundation of China (61175003), Hundred Talents Program of CAS, The strategic Priority Research Program of CAS (XDA06030300), and National Basic Research Program of China (2012CB316300).

References

1. Giannakopoulos, T., Kosmopoulos, D.I., Aristidou, A., Theodoridis, S.: Violence Content Classification Using Audio Features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
2. Giannakopoulos, T., Pikrakis, A., Theodoridis, S.: A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks. In: Proceedings of the 9th International Workshop on Multimedia Signal Processing, pp. 90–93. IEEE Press, Crete (2007)
3. Gong, Y., Wang, W.-Q., Jiang, S., Huang, Q., Gao, W.: Detecting Violent Scenes in Movies by Auditory and Visual Cues. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 317–326. Springer, Heidelberg (2008)
4. Lin, J., Wang, W.: Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 930–935. Springer, Heidelberg (2009)
5. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-Visual Fusion for Detecting Violent Scenes in Videos. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 91–100. Springer, Heidelberg (2010)
6. Wiskott, L., Sejnowski, T.: Slow Feature Analysis: Unsupervised Learning of Invariances. Neural Computation 14, 715–770 (2002)
7. Zhang, Z., Tao, D.: Slow Feature Analysis for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 436–450 (2012)
8. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of Computer Vision and Pattern Recognition, pp. 3169–3176. IEEE Press, Providence (2011)
9. Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003)
10. Chang, C.C., Lin, C.J.: LIBSVM:A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. 27, 27:1–27:27 (2001)
11. Laptev, I., Lindeberg, T.: Space-time Interest Points. In: Proceedings of International Conference on Computer Vision, pp. 432–439. IEEE Press, Nice (2003)