Image and Vision Computing 48–49 (2016) 37–41

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Violence detection using Oriented VIolent Flows[☆,☆☆]

Yuan Gao[a], Hong Liu[a,*], Xiaohu Sun[a], Can Wang[a], Yi Liu[a,b]

[a] *Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China*
[b] *IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institution, Shenzhen 518057, China*

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 13 January 2016
Accepted 14 January 2016
Available online 24 February 2016

Keywords:
Violence detection
Oriented VIolent Flows
AdaBoost
SVM

ABSTRACT

Nowadays, with so many surveillance cameras having been installed, the market demand for intelligent violence detection is continuously growing, while it is still a challenging topic in research area. Therefore, we attempt to make some improvements of existing violence detectors. The primary contributions of this paper are two-fold. Firstly, a novel feature extraction method named Oriented VIolent Flows (OViF), which takes full advantage of the motion magnitude change information in statistical motion orientations, is proposed for practical violence detection in videos. The comparison of OViF and baseline approaches on two public databases demonstrates the efficiency of the proposed method. Secondly, feature combination and multi-classifier combination strategies are adopted and excellent results are obtained. Experimental results show that using combined features with AdaBoost+Linear-SVM achieves improved performance over the state-of-the-art on the Violent-Flows benchmark.

1. Introduction

Violence detection is a particular problem within a greater problem of action recognition. In the last years, automation recognition of human actions in realistic videos has become increasingly important for applications such as video surveillance, human–computer interaction and content-based video retrieval [1,2]. Recent proposed methods for action recognition can be roughly grouped into local, interest-point based, or global, frame-based methods.

In local methods, spatio-temporal feature points [3] are detected to represent human activity in a video [4]. An unsupervised method similar to the bag-of-words approach is proposed to learn the probability of distribution of these feature points [5]. Then a video can be represented with bag-of-feature techniques [6]. However, when there are only few interest points or too much motion, they may fail to provide enough meaningful information. On the other hand, global methods use global features such as optical flow to represent the state of motion in the frame at a particular instant of time [7,8]. In [7], optical flow histograms, based on horizontal and vertical directions, were used as action descriptors to address the problem of human action recognition. Wang et al. in [8] employed optical flow to obtain densely tracking sampled points, and then they utilized these dense trajectories to calculate local descriptors for action recognition. Optical flow can describe coherent motion of moving objects, which is a good feature for motion detection and tracking. As a consequence, it has been widely used for object tracking and motion representation.

1.1. Relation to prior work

For violence detection, there existed some studies utilizing both vision and acoustic technologies [9,10,11]. However, in lots of surveillance systems, audio cues are usually unavailable. Therefore, in this paper we focus on violence detection in videos using pure vision methods. Datta et al. made an early attempt to address the violence detection problem based on background substraction [12]. Nevertheless, for violence happening in crowded scenarios, this approach may fail. In Refs. [13,14], the presence or absence of blood is an important cue for violence recognition. However, when the surveillance cameras only output gray-scale videos, the performances of these approaches could be affected. More recently, Clarin et al. used local interest-point based approaches to detect fights on their own designed dataset [15]. Nievas et al. proposed a novel descriptor called ViF for real-time crowd violence detection [16]. Two databases, Hockey Fights and Violent-Flows, which were proposed in [15,16] separately, are benchmarks in our experiments for

[☆] This paper has been recommended for acceptance by Nicu Sebe.

[☆☆] This work is supported by National Natural Science Foundation of China (NSFC, No.61340046), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Science and Technology Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, No. JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province (No.2015A030311034).

[*] Corresponding author.
E-mail addresses: ygao@sz.pku.edu.cn (Y. Gao), hongliu@pku.edu.cn (H. Liu), xiaohusun@pku.edu.cn (X. Sun), canwang@pku.edu.cn (C. Wang), yi.liu@imsl.org.cn (Y. Liu).

http://dx.doi.org/10.1016/j.imavis.2016.01.006
0262-8856/© 2016 Elsevier B.V. All rights reserved.

that both of them contain real-world, unconstrained, and violent or non-violent videos. Deniz et al. applied the Radon transform to get extreme acceleration patterns, which are the main feature of their method [17]. After using AdaBoost as the classifier, the violence recognition rate improved compared with BoW(STIP) and BoW(MoSIFT) methods on the Hockey Fights dataset. In 2015, Rota et al. used the improved trajectories [18] to get a feature codebook and the interpersonal space to detect violent interaction [19]. The disadvantage of their method is that it heavily relies on an accurate pedestrian detector or tracker and only analyzes the circumstance of the interaction between two people, which could hardly be applied to analyze videos containing turbulent crowds in the Violent-Flows dataset.

In this paper, a new feature, OViF, is proposed for violence detection. The original motivation of designing OViF is to make full use of the orientation information of optical flow, which is omitted by ViF. Moreover, since feature combination and multi-classifier combination are common manners in the classification process, they are also adopted by us in experiments. During the period of multi-classier combination, a Linear-SVM classifier, which is trained on the features selected by AdaBoost, achieves noticeable improvement in violence recognition rate. This indicates the advantage of using AdaBoost as a feature selector. Finally, the combined features, ViF + OViF, obtain the state-of-the-art violence detection performance when using AdaBoost + Linear-SVM, which also shows the effectiveness of our proposed OViF features.

2. Feature extraction

In this section, two kinds of feature extraction methods, ViF and our proposed OViF, will be introduced orderly.

2.1. VIolent Flows (ViF)

The ViF descriptor is initially designed for crowd violence detection in [16]. In order to get a ViF vector for a video sequence, there are three steps. Firstly, through computing optical flow between pairs of consecutive frames, the magnitude of the flow vector, which corresponds to any pixel in a frame, can be computed. Then, by comparing these motion magnitudes between sequential frames, the magnitude-change maps are calculated, which helps obtaining a mean-magnitude map. Lastly, the obtained mean-magnitude map is divided into $M \times N$ non-overlapping regions, in each of which the frequencies of magnitude changes are collected and represented as a fixed-size histogram. And the final ViF vector is the concatenation of these histograms.

2.2. Oriented VIolent Flows (OViF)

As introduced above, ViF is a feature descriptor which can describe the changes of observed motion magnitudes well. However, in some cases, it may loss some important information. For example, when the flow vectors of the same pixel in two sequential frames have the same magnitude but only differ in their directions, the effect of ViF seems to be restricted. This is because that ViF thinks that there is no difference between these two flow vectors but actually they differ a lot. Therefore, we propose a new feature representation method, OViF, which depicts the information involving both of motion magnitudes and motion orientations. The visualization of extracting OViF for a video sequence is illustrated in Fig. 1. And the details of this process is introduced as below:

Firstly, the optical flow should be calculated between pairs of sequential frames in the input video sequence. The flow vector of each pixel can be represented as:

$$\left|V_{i,\,j,\,t}\right| = \sqrt{\left(V^{x}_{i,\,j,\,t}\right)^2 + \left(V^{y}_{i,\,j,\,t}\right)^2} \tag{1}$$

$$\Phi_{i,\,j,\,t} = \arctan\left(V^{y}_{i,\,j,\,t}/V^{x}_{i,\,j,\,t}\right). \tag{2}$$

Here, t means the t-th frame in a video sequence, and (i, j) indicates the pixel location. Then, for each flow map which corresponds to a frame, we partition it into $M \times N$ non-overlapping blocks. After this, since 360° can be equally divided into B sectors and each sector corresponds to a bin of a histogram, the flow-vector magnitude $|V_{i,j,t}|$ is added into the bin where the flow-vector angle $\Phi_{i,j,t}$ locates. It is clear that, for each block, we get a histogram. These histograms are then concatenated into a single vector H, which is called Histogram of Oriented Optical Flow (HOOF) with X-dimensions:

$$X = M \times N \times B. \tag{3}$$

This calculation step is exhibited in Step 3 of Fig. 1. Note that the HOOF here is special designed for violence detection task and is a little different from that in [20]. There is no normalization here and the ways of counting orientations also differ. Subsequently, the HOOF vectors are used to obtain binary indicators:

$$b_{x,\,t} = \begin{cases} 1 & \textit{if } \left|H_{x,\,t} - H_{x,\,t-1}\right| \ge \theta. \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

Here, x is the x-th dimension of the feature vector H, and θ is the average value of $|H_{x,t} - H_{x,t-1}|, x \subseteq [1, X]$. The above equation explicitly reflects the magnitude changes in different sectors. The mean magnitude-change vector is donated as:

$$\overline{b_x} = \frac{1}{T}\sum_t b_{x,\,t}. \tag{5}$$

Finally, $\bar{b}$ is the final OViF vector for a sequence of frames, which counts the motion magnitude change frequencies in both direction sectors and spatial regions. Another reason for us to design OViF is that, based on observations of violent and nonviolent videos, we find that movements in nonviolent videos usually have the same direction with little deviation, while in violent videos movements are disordered with large deviation. Consequently, we encode the orientation information of motion and propose OViF.

3. Classifiers

Two types of traditional and popular machine learning algorithms, SVM and AdaBoost, are utilized in this work.

- As for the first classifier, a linear SVM [21] is selected in consideration of its simplicity, effectiveness and, last but not least, the speed.
- Regarding the second classifier, Gentle AdaBoost [22] is chosen for that it is one of the most practically efficient boosting algorithms.
- Apart of using these two algorithms individually, the combination of AdaBoost and SVM is also an effective way to improve classification performance. In particular, AdaBoost is only applied to select features and then a SVM classifier is trained on the selected features.

To know the specified implementation of these two algorithms, readers can refer to LIBLINEAR[1] and GML AdaBoost Matlab Toolbox[2] .

[1] http://www.csie.ntu.edu.tw/cjlin/liblinear/.

[2] http://graphics.cs.msu.ru/ru/science/research/machinelearning/adaboosttoolbox/.