Optik 127 (2016) 795–801

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Recognizing violent activity without decoding video streams

Jianbin Xie [a], Wei Yan [a], Chundi Mu [a], Tong Liu [a,*], Peiqin Li [a], Shuicheng Yan [b]

[a] National University of Defense Technology, College of Electronic Science and Engineering, Kaifu District, Yan W Pond #47, Changsha 410073, Hunan, China
[b] National University of Singapore, Singapore, Singapore

ARTICLE INFO

Article history:
Received 26 January 2015
Accepted 27 October 2015

Keywords:
Activity recognition
Violent activity
Motion vectors

ABSTRACT

The processes of motion target detection and tracking in most of traditional activity recognition methods are usually complicated and the application of these methods is limited. In this paper, we propose a fast violent activity recognition method based on motion vectors. First, we extract the motion vectors from compressed video data directly. Then, we analyze the features of the motion vectors in each frame and between frames, and get Region Motion Vectors descriptor (RMV). Finally, we use the Support Vector Machine (SVM) which takes the radial basis as the kernel function to classify the RMV and determine whether the violent activity exists in the video or not. Experimental results on several datasets have shown that the proposed method can detect 96.1% of the violent activities in videos (false probability is about 5.1%), and the calculation speed is very fast, which means the new method can be used in embedded systems.

1. Introduction

Violent activity is a kind of very harmful action for people and society, which takes the human body or property as the target, and uses violent means to endanger people's life, health and personal freedom. Violent activity recognition based on video analysis refers to analyzing the motion feature of the target in the videos to determine whether the violent activity exists in the video or not. In a video surveillance system, the earlier the violent activity is detected, the lower the harm will be.

Violent activity is not a kind of fixed activity (e.g., raising hand and stooping), or a simple activity (e.g., walking and riding bike), but a kind of complicated space–time interactive activity, which does not have fixed model or style, and is very hard to be defined and recognized exactly.

Existing video analyzing algorithms include target detection and target recognition. The ways of separating the target and background, tracking the target and extracting the key points require expensive computation. There are some limitations when those algorithms are used in real applications. For example, most of the video capture and coding devices do not have enough resource for real-time operation of those algorithms, and software and hardware, which use those algorithms to analyze the huge amount of video recording data, are usually too expensive to be widely used.

In video monitoring systems, many algorithms have been proposed to reduce the transmission bandwidth, such as motion detection and motion compensation. Therefore, when we get compressed video data, the motion vectors of the video can be got without additional calculation. The motion vectors represent the relative motion between the current macro-blocks and reference macro-blocks, which are not exactly equivalent to the actual movements of the target. However, we can still get some useful information from them. By analyzing the implications of the motion vectors, complexity and processing time of the algorithm can be reduced significantly, and the demand for the hardware can also be reduced.

To facilitate the violent activity recognition study, we firstly collect the positive video clips (e.g., boxing and fighting) and negative video clips (e.g., walking and jumping) from public video datasets such as UCF sports, UCF50, HMDB51 and YouTube, etc., to construct the Video streams for Violent Activity Recognition dataset v1.0 (VVAR10) for violent activity analysis. Then, we extract the motion vectors of the video clip in VVAR10, analyze the features of the motion vectors and use Region Motion Vectors descriptor (RMV) to describe the motion features of the video clip. Finally, we use the Support Vector Machine (SVM) to determine whether the violent activity exists in the video or not by machine learning. Eventually, the extensive experiments on the VVAR10 demonstrate the average precision of our method is about 96.3% and the false probability is about 5%, and the calculation speed is very fast, which means the method can be implemented in embedded systems.

* Corresponding author. Tel.: +86 13574870542.
E-mail addresses: jbxie@126.com (J. Xie), 15660565@qq.com (W. Yan), 723840642@qq.com (C. Mu), liutong1129@126.com (T. Liu), lipeiqin_nudt@163.com (P. Li), eleyans@nus.edu.sg (S. Yan).

http://dx.doi.org/10.1016/j.ijleo.2015.10.165
0030-4026/

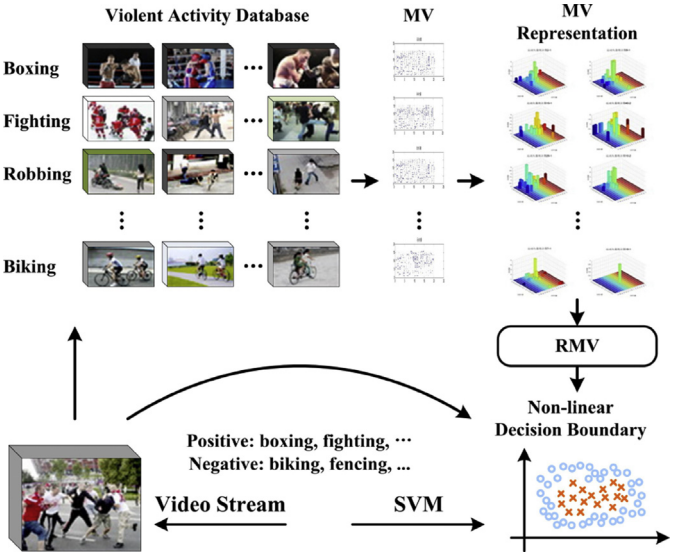

Fig. 1. We select several video clips which contain four types of violent activities and lots of non-violent video clips to form VVAR10 dataset (Section 3). Then we use Region Motion Vectors descriptor to describe the feature of motion vectors of videos and use SVM to classify them (Section 4). Motion vectors can easily be gotten from compressed video data, and SVM is very suitable for dealing with non-linear classification requirement.

Fig. 1 illustrates the proposed framework for violent activity recognition by motion vectors. The main contributions of this work can be summarized as follows:

- We propose a new method for violent activity recognition, which has less calculation and can be widely used in smart video surveillance systems.
- The proposed method can take full advantage of the motion vectors existing in compressed video data and reduce the time of calculating the motion vectors.
- The proposed method can use RMV to describe the motion features of the video data.

The rest of the paper is organized as follows. Section 2 discusses the related work. Then, we describe the construction of VVAR10, and propose the framework for violent activity recognition in Sections 3 and 4, respectively. Experiments and discussions are presented in Section 5. Section 6 concludes this work.

2. Related work

Violent activity recognition is an important part of human activity recognition due to its potential for a multitude of applications [10]. As we all know, "fighting between two persons" is an interaction between two persons [1]. Thus, the analysis of interactions is to analyze the activities between persons. For activity recognition, early works focus on classifying video sequences of a single person in controlled environments. In this scenario, the background is simple and uniform [3,11,12]. However, with the development of activity recognition technologies, researchers have tried to introduce more natural and unconstrained videos. For instance, Laptev et al. have studied the sequences from feature films [13], and some researchers have focused on the recognition of "wild videos", such as the video from YouTube [14,18]. The approaches of action recognition are varied, and so is the way to classify them [1,16,19].

There are some approaches that can analyze the video directly. Messing et al. presented an activity recognition feature inspired by human psychophysical performance which is based on the velocity history of tracked key points [2]. Schuldt et al. constructed video representations in terms of local space–time features and integrated such representations with SVM classification schemes for recognition [3]. Wang et al. introduced a novel descriptor based on motion boundary histograms, which have shown good performance by classifying many kinds of actions [5]. Sadanand et al. presented Action Bank, a new high-level representation of the video, which is comprised of many individual action detectors sampled broadly in the semantic space as well as the viewpoint space [7]. Ryoo et al. can recognize human high-level activities such as "fight" and "assault" by defining complex human activities based on simpler activities or movements by context-free grammars [8]. Jingen Liu et al. used high-level semantic concepts to realize the recognition of human actions. Sun and Nevatia used fisher vectors to realize the classification of large scale web video event [32].

With the development of the sensor industry, sensors are widely used in computer version [15]. Some researches recognize the activity with the help of the data got by sensors. Ravi et al. used a triaxial accelerometer worn near the pelvic region as a motion detector to recognize activities [4]. Ravi et al. reported their efforts to recognize user activities from accelerometer data [4]. Spriggs et al. explored first-person sensing through a wearable camera and Inertial Measurement Units [9]. Maekawa et al. introduced a way to recognize actions with sensors on the wrist [17].

In summary, we find that previous studies always focus on the recognition of the basic action, and in order to get the accurate space–time features and trajectories, many of them need the help of accessories worn on the body, such as accelerometer and GPS. Hierarchical approaches are used for the recognition of complex activities, such as Markov Network [6,22,29], Conditional Random Fields [21], and so on [20,28,31]. It is assured that these approaches are good. However, due to their complexity, they are not suitable for real-time monitoring and warning systems. Xie et al. proposed a fast and robust algorithm for fighting behavior detection based on motion vectors, but the motion vectors were extracted by an improved Three-Step Search method [33]. In this paper, we attempt to recognize the violent activity by analyzing motion vectors, which can be extracted directly from the video stream, and the way to analyze the method will be described in Section 4.

3. VVAR10 dataset

There are several datasets, e.g., Weizmann [9], KTH [10], UCF sports [1] and Hollywood2 actions [11] for activity recognition. However, none of them is directly suitable as they usually focus on the recognition of simple individual actions. Therefore, to study our proposed problem, we need a large dataset of violent actions, which contains fighting, boxing, hammering and pursuing.

Hence, we have built VVAR10 dataset which contains 296 positive samples and 277 negative samples. We get positive samples and negative samples from the UCF sports, the UCF50, the HMDB51 [7] and the YouTube [27]. In order to save the experimental time and test the effectiveness of our algorithm, we partition each video clips less than five seconds. To diversify our dataset, we select violent activity videos of various situations, day and night, single and multiplayer, with tools and non-tools, and so on. Fig. 2 shows some examples in VVAR10.

4. The proposed framework

To determine whether the violent activity exists in the video or not, two key issues are required to be addressed. First, most of the video coding algorithms use the macro-block as the basic process unit, and often use deformable macro-block technology (for example, there are 7 different sizes of macro-blocks in H.264) and multiple reference frame technology to improve the coding efficiency, so the motion vectors in the same frame often have