Violence Detection With Two-Stream Neural Network Based on C3D

zanzan Lu, Minnan Normal University, China
Xuewen Xia, Minnan Normal University, China
https://orcid.org/0000-0002-4938-1479
Hongrun Wu, Minnan Normal University, China
Chen Yang, School of Artificial Intelligence, Shenzhen Polytechnic, China

ABSTRACT

In recent years, violence detection has gradually turned into an important research area in computer vision with many proposed models with high accuracy. However, there is unsatisfactory generalization ability of these methods over different datasets. In this paper, the authors propose a violence detection method based on C3D two-stream network for spatiotemporal features. First, the authors preprocess the video data of RGB stream and optical stream respectively. Second, the authors feed the data into two C3D networks to extract features from the RGB flow and the optical flow respectively. Third, the authors fuse the features extracted by the two networks to obtain a final prediction result. To testify to the performance of the proposed model, four different datasets (two public datasets and two self-built datasets) are selected in this paper. The experimental results show that the model has good generalization ability compared to state-of-the-art methods since it not only has good ability on large-scale datasets but also performs well on small-scale datasets.

KEYWORDS

C3D Network, Deep Learning, Feature Fusion, Spatiotemporal Features, Two-Stream Network, Video Detection, Violence Detection, Violent Datasets

INTRODUCTION

With the development and progress of society, a harmonious and stable social security becomes crucial. Thus, how to timely detect violence and then effectively reduce or prevent violent crimes become very important in now a days. Recently, various methods for feature extraction on images have been developed for a long time benefited from the rapid progress of computer technology, so extraction on images is relatively mature. For instance, in (LeCun, Bottou, Bengio, 1998), LeNet was firstly applied in handwritten number recognition, and ResNet (He, Zhang, et al, 2016) has recently gained popularity for image recognition. The accuracy of some image recognition methods has even surpassed that of humans. For example, Russakovsky et al. (Russakovsky, Deng, et al, 2015) evaluated human classification error is 5.1% on a large-scale image dataset (ILSVRC2012-2014 classification test set), while the error of the ResNet model mentioned above was only 3.6% on the same dataset.

Unlike traditional image recognition, the violence detection, as a subclass of action detection, often occurs over a continuous period of time. Hence, an algorithm applied in the violence detection must take into account not only the spatial dimension of the image, but also the temporal dimension

DOI: 10.4018/IJCINI.287601

of the image. This makes it necessary for researchers to make their models capable of extracting spatiotemporal features simultaneously, as is the case in this paper. Based on the property, a number of models have been proposed in the last few decades (Ramzan, Abid, et al, 2019), such as two-stream network (Simonyan, Zisserman, 2014) and 3D Convolutional network (C3D) (Tran, Bourdev, et al, 2015; Ji, Xu, et al, 2012), etc. Each of these methods extracts the temporal and spatial information of the image in their own way, and yields its own characteristics.

In order to apply violence detection to real-world applications, it is necessary that the violence detection model has a favorable generalization capability. However, the generalization ability of existing models is still a challenge. For instance, unlike traditional machine learning, in order to obtain a promising generalization ability, majority of deep learning models need a large amount of data for training (Simonyan, Zisserman, 2014; Tran, Bourdev, et al, 2015), and the reality is that sometimes there is not that much data, which leads to unsatisfactory performance of the model. In order to solve this problem, researchers not only need to improve the model, but also to optimize the datasets.

Violence detection, as a subclass of action recognition, can use many of the methods of behavior recognition. Two-stream network detect action by combining temporal and spatial streams, an approach that has the advantage of being easily scalable and highly accurate. The C3D network uses the convolution of 3D structures for action recognition, which has the characteristics of structural simplicity and efficiency. Inspired by the excellent performance of two-stream network and C3D network in the field of action recognition, the authors tackle the challenges mentioned above by proposing a C3D-based two-stream network violence detection model.

The main contributions of this paper include:

1. Current mainstream methods are unable to learn effective models due to the lack of data on violence. Moreover, the existing violence detection datasets suffer from the problem of insufficient data quantity and single data scene. To tackle these problems, the authors collect violent videos from websites and process them. In this way, the authors create two new datasets in this work, i.e., public datasets and self-built datasets.
2. The authors find that previous models (e.g., C3D) did not perform very well on small-scale datasets, which indicates the low generalization ability of previous models on small-scale datasets. To further improve the generalization ability of the model, the authors improve the two-stream network so that authors' C3D two-stream network can not only perform well on large-scale datasets, but also achieve high accuracy on small-scale datasets.

The remaining parts of the paper is organized as the following. Section briefly describes related work in the area of violence detection. The detail of the proposed model is introduced in Section 3, while the detail of two types of datasets are described in section 4. Section 5 contains the results and discussion of the experiment and the conclusion and future work are presented in section 6.

Related Work

Violence detection, as a hot topic in computer vision applications, is increasingly sought after by researchers. Today, many techniques have emerged in the field of violence detection, and some have even been applied in the real world (Aktı, Tataroğlu, Ekenel, 2019; Febin, Jayasree, Joy, 2019; Roshan, Srivathsan, et al, 2020; Singh, Patil, Omkar, 2018). Generally, these techniques fall into two main categories: the traditional machine learning and the recently proposed deep learning.

In the early days, researchers generally used the traditional machine learning methods to extract features of violent videos. For instance, Clarin et al. (Clarin, Dionisio, et al, 2005) suggested a motion intensity analysis on skin and blood to detect movie violence. Similarly, Chen et al. (Chen, Hsu, et al, 2011) integrated facial, blood and movement information to determine if the action scene has violent content. The main contribution of the work is applying the presence of highly relevant objects