IOP Conf. Series: Journal of Physics: Conf. Series **1487** (2020) 012033 doi:10.1088/1742-6596/1487/1/012033

An Action Recognition Method Based on Deformable Convolution Network

Xu Dong, Li Tan*, Lina Zhou and Yanyan Song

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, China

*Email: tanli@th.btbu.edu.cn

Abstract. In order to solve the problem of behavior recognition in short video, this paper first proposes a key frame extraction algorithm based on mutual information entropy, which uses sliding window to preserve the timing information between frames. Based on the key frame extraction, a based on Deform-GoogLeNet, a two-stream CNN method for deformable convolutional networks, uses the two-stream network to extract the RGB features and optical Flow characteristics of the image separately, and uses the weighted average method to obtain the results of behavior recognition. On the public dataset Charades dataset, mAP is 22.9, which is higher than the similar fusion algorithm, which proves that the proposed algorithm has a good effect in short video behavior recognition.

1. Introduction

With the development of communications technology and the widespread use of mobile devices, the information carriers have a tendency to change from texts and pictures to videos. According to the reports, the total used time in short video platforms has achieved 5.1 billion hours in February,2018 and has been growing continuously. The advantages of short video are obvious, though, the disadvantages are also obvious [1,2]. Firstly, some short videos are fragmented and the content of them are not complete enough, which is easy to be taken out of context and even turned into video rumors, creating a bad influence to society. Secondly, some short videos content is shoddy and vulgar which have a bad impact on young people. Therefore, to make up for the shortcomings of short videos and to make the content of short videos healthier, it is necessary to learn the content of short videos effectively.

The information of human actions is the key to learning, which helps people to solve the goal of "what is he or she doing". In Baidu Encyclopaedia, short videos refer to the videos, length of which is within 1 minutes and make it difficult to understand the actions in the videos. Besides, if the redundant frames in the short videos are not processed, the information cannot be accurate and a lot of storage resources are occupied. Therefore, before extracting the action features, it is necessary to extract the key frame [3,4] and after that, the action features are extracted based on the key frames so that more accurate behavior characteristics can be obtained.

In computer vision, deep learning methods do well in image processing than the traditional algorithms. The advantages of deep learning are mainly manifested in two aspects: Firstly, deep learning methods perform well in recognition and classification [5,6]; Secondly, deep learning network can be used in different application scenarios by fine-tune methods [7,8], which can get better

performance models. So, it is feasible to use deep learning methods to build models and understand the short videos [2,9,10]. The main contributions of this paper are as follows:

1) Considering the temporal relationship between frames, a key frame extraction method based on mutual information entropy is proposed to extract the key frames, preserving the timing information of the video as much as possible.

2) After the key frame extraction, an action recognition method based on deformable convolutional network is proposed to improve the efficiency of action recognition.

2. Related work

2.1 Key frame extraction

Early key frame extraction was mainly processed from the perspective of lens detection. Vilia[11] et al. proposed a boundary detection algorithm that uses image color changes and thresholds to compare, but still contains more redundant information; Hannane et al.[12] A key frame extraction method based on SIFT has certain practicability under different lighting conditions, but key frame extraction from the perspective of lens detection can't complete the presentation of video content; Zhuang et al.[13] use K-Means algorithm to frame As a sample point, the key frame is located in the cluster center, but the K-Means algorithm is affected by the initial value, so it does not reflect the key frame information well; Hu et al. use the optical Flow difference method to perform key frames in the video. The extraction has achieved a certain effect.

2.2 Action Recognition

The behavior recognition based on deep learning is mainly divided into two categories, namely spatio-temporal network and two-stream network. The spatio-temporal network focuses on extracting temporal information in the video. The methods mainly include CNN+LSTM[14,15] and C3D methods[16,17]. For example, the CNN+LSTM method first extracts the spatial features in the image through the CNN network, and then uses LSTM and other methods to extract the time information in the video, similar to the serial structure in the circuit. In 2013, Ji et al. [18] proposed the C3D method and they used traditional CNN network processed 2D image features and added a dimension information based on this, using a CNN network approximation as a time dimension to predict the possible next action. In 2014, Simonyan et al. [19] proposed the method of Two stream CNN. They used two convolutional neural networks and divided the video image information into RGB feature information and optical Flow feature information. The RGB features use single frame, and the optical Flow features use continuous frames. Finally, the features obtained were identified by SVM, and had a good performance.

3. Our methods

3.1 Key frame extraction

Before feature extraction, key frame extraction is required to minimize redundant frames to save storage space. The selection of key frames should be representative and be as dissimilar as possible so that the key frames have more information. In this section, a method for extracting key frames based on mutual information entropy will be proposed.

3.1.1 Mutual information entropy. On the basis of information theory, Viola and Collignon proposed the concept of mutual information. For images A, B, mutual information is defined as:

$$I(A,B) = H(A)+H(B)-H(A,B) \tag{1}$$

where $H(A)$ and $H(B)$ are the information entropy of the image and have the following definitions.

$$H(A) = -\sum_{\alpha} P_A(a)\log P_A(a) \tag{2}$$

$$H(B) = -\sum_{\beta} P_B(b)\log P_B(b) \tag{3}$$

$H(A,B)$ is joint information entropy, defined as follow:

$$H(A,B) = -\sum_{\alpha,\beta} P_{a,b}(a,b)\log P_{ab}(a,b) \tag{4}$$

where P_A, P_B are the probability distributions of images A and B, and P_{AB} is the joint probability distribution. The larger the $I(A,B)$, the greater the correlation between A and B, and vice versa, the less relevant A and B are.

3.1.2 Key frame extraction algorithm based on mutual information entropy. Due to the sequence relationship between frames in video, our algorithm is as follows:

1) Framing the short video to get the original frame sequence $\{f_0, f_1, f_2, f_3, \cdots, f_n\}$;

2) Set frame 0 as the initial key frame, $f_{k_0} = f_0$, put it in keyframe candidate sequence $Candidates\{f_{k_0}\}$ and set the sliding window value ω;

3) Traversing the sequence of frames from frame 1. In the frame sequence, the value of the sliding window is the i-th frame to the $i+\omega$ th frame, $i<n$ and calculating the mutual information entropy $I(f_i, f_{k_i})$ of f_{k_i} in each frame and key frame candidate sequence in the sliding window, which is

$$I(f_n, f_{k_j}) = H(f_n) + H(f_{k_j}) - H(f_n, f_{k_j}),\ n \in (i, i+\omega) \tag{5}$$

where $I(f_n, f_{k_j})$ is the mutual information entropy of the n-th frame and the k_j candidate frame, $n \in (i, i+\omega)$ is the sliding time window range, f_{k_j} is the frame added in the candidate frame sequence;

4) According to the concept of mutual information entropy introduced in the previous section, the smaller the mutual information entropy value, the less relevant the two images are. Therefore, find the minimum value of the mutual information entropy in the $(i, i+\omega)$, denoted as $\min(I(f_n, f_{k_j}))$, and the position of the corresponding frame, denoted as $F_{candidate}$;

5) Add $F_{candidate}$ to the key frame candidate sequence $Candidates\{f_{k_0}\}$; and move the sliding window, repeat steps 3), 4) until the sliding window ends to slide, to obtain a complete key frame sequence;

3.2 Action Recognition Based on Deformable Convolution Network

3.2.1 Deformable convolution network. The traditional convolution is sampling in the specification point and the kernels are generally fixed convolution kernel such as 3×3, 5×5 and 7×7, etc. Although the feature point can be extracted, the square convolution kernel is not necessarily the most suitable [20]. The convolution kernel does not accommodate sampling points of different shapes. The emergence of deformable convolution networks can adapt to deformable sampling points. The example of the deformable kernel is shown in Figure 1.

(a) (b)

Figure 1. (a) is the sampling structure of the traditional 3x3 convolution kernel; (b) deformable convolution kernel

Traditional convolution kernels need to directly train the parameters of the convolution window, while deformable convolution networks require an additional convolutional network to train the shape of the convolution window to get the offset of the pixel. For traditional convolution output:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{6}$$

The opponent's grid is deformed and offset which need to add Δp:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{7}$$

3.2.2 Two-stream CNN network structure. In this section, our work is to combine the characteristics of two-stream network and deformable convolution network to propose a two-stream network structure of Deformable-GoogLeNet for action recognition, the work Flow is as Figure 2.

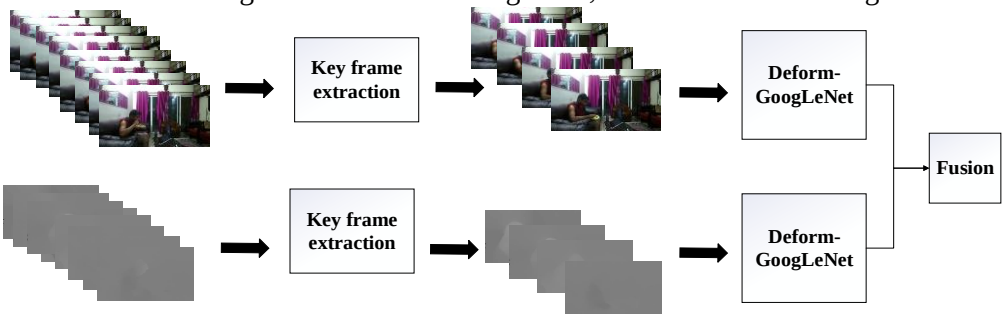

Figure 2. Two-stream DeformGoogLeNet

The fusion method is that the RGB features and optical Flow features acquired by the deformable convolution network, need to be weighted and fused to obtain the final probability result and the result with high probability is selected as the final result. The following formula is

$$\text{Fusion} = \alpha \times \frac{F_{rgb}}{n} + \beta \times \frac{F_{flow}}{n} \tag{8}$$

where $\alpha+\beta=1$, F_{rgb} is the extracted image RGB features, and F_{flow} is the extracted image optical Flow features.

4. Experimental results

4.1 Experimental platform

The operating system used in this experiment is Ubuntu16.04, CPU: Intel(R) Core i5 -7300HQ, GPU: NVIDIA k40c×2, Cuda version: cuda 8.0, deep learning framework: PyTorch.

4.2 Experimental parameter setting

4.2.1 Key frame extraction settings. To verify the effectiveness of the proposed method, three short videos were selected for experimentation on the Charades dataset and 004QE. Because the definition of key frames is not uniformly defined, the video key frame set is constructed by manually marking the key frames in the experiment for testing. We use two evaluation indicators which are Precision and Recall. Among them, the Precision is

$$Precision = \frac{n_{true}}{n_{extract}} \tag{9}$$

where n_{true} is the number of frames matched between the manually labelled keyframes and the keyframes extracted by our method, $n_{extract}$ is the number of key frames extracted by our method. And the Recall is

$$Recall = \frac{n_{true}}{n_{annotation}} \tag{10}$$

where $n_{annotation}$ is the number of key frames manually labelled.

4.2.2 Action recognition settings. The data set selected in the experiment is Charades, and the data set has a total of 9848 video segments, of which there are 157 behavior categories in the data set. The training set data is 7985 video segments, and the test set is 1863 video segments. Each video length is 30.1s. The Deformable-GoogLeNet training process uses a backpropagation method with Adam optimization, batchsize is set to 32, momentum is set to 0.9, and base learning rate is set to 0.01. The evaluation index is the mean average precision (mAP).

4.3 Results

The results of the key frame extraction algorithms at 004QE are shown in Table 1**,** it can be seen that in the process of extracting short video key frames, precision tends to increase with the length of time window, while recall tends to decrease with the increase of time window. The reason may be that the number of key frames to be extracted is small, resulting in lower recall and higher precision values.

Table 1. The results of the key frame extraction algorithms in charades

Name	**Evaluations**						
	window length	**Overlap window**	**Total number of frames**	**Total number of key frames**	**Compression ratio**	**Precision**	**Recall**
004QE	4	2	735	368	0.501	0.511	0.905
	8	4	735	184	0.250	0.707	0.833
	12	6	735	123	0.167	0.780	0.615
	16	8	735	92	0.125	0.913	0.538
	20	10	735	74	0.101	0.838	0.397

After key frame extraction and deep network feature extraction in RGB, the following results are shown in Table 2. From Table 2, it can be seen that the Deformable-GoogLeNet proposed in our work has a good performance in RGB feature extraction. TrainPrec1 is 20.925, TrainPrec5 is 62.149 and mAP is higher than the other method which means our method can effectively extract RGB features of images.

Table 2. Comparison results of other algorithm in RGB feature extraction in Charades

Methods	**Results**				
	TrainPrec1	TrainPrec5	ValPrec1	ValPrec5	mAP
ResNet34	18.933	55.220	6.394	21.391	0.165
Our methods	**20.925**	**62.149**	**8.715**	**26.947**	**0.254**

Finally, RGB and FLOW features are fused by weighted fusion method, and the results are shown in Table 3. It can be seen that the mAP of the fusion method proposed in this paper is 22.9, which is better than the two stream method. The result proves the effectiveness of the proposed algorithm.

Table 3. Comparison results of two stream algorithm in RGB feature extraction in Charades

methods	**Results**	
	modality	mAP
2-Stream	RGB+Flow	18.6
Our methods	**RGB+Flow**	**22.9**

5. Conclusion and future work

In this paper, a key frame extraction algorithm based on mutual information entropy is proposed. On the basis of key frame extraction, a Deformable-GoogLeNet method is proposed to solve action

recognition, which further improves the performance of action recognition in two-stream CNN network. Through experiments on Charades, the mAP value of this method is 22.9, which is higher than similar two-stream CNN algorithms. The effectiveness of the proposed method is proved.

Acknowledgments
This research was funded by the National Natural Science Foundation of China grant number (61702020), Beijing Natural Science Foundation grant number (4172013) and Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund grant number (L182007).

References

[1] Zhang, J. , Nie, L. , Wang, X. , He, X. , & Chua, T. S. . 2016. *Shorter-is-Better: Venue Category Estimation from Micro-Video.* the 2016 ACM. ACM.
[2] Meng Liu,Liqiang Nie,Meng Wang and Baoquan Chen, 2017 Towards Micro-video Understanding by Joint Sequential-Sparse Modeling, the 2017 MM.
[3] Zong, Z. , & Gong, Q. .2017. Key frame extraction based on dynamic color histogram and fast wavelet histogram.2017 IEEE International Conference on Information and Automation (ICIA). IEEE.
[4] Meng Jian, Shijie Zhang, Xiangdong Wang,etc. 2018, Deep Key Frame Extraction for Sport Training, Neurocomputing, (September. 2018)
[5] Li, H. 2017. Multi-scale Spatial Topic Models for scene recognition.International Congress on Image & Signal Processing. IEEE.
[6] Chaudhari, P. , & Agarwal, H. . (2017). Progressive Review Towards Deep Learning Techniques. Proceedings of the International Conference on Data Engineering and Communication Technology. Springer Singapore.
[7] Zhou, Z. , Shin, J. , Zhang, L. , Gurudu, S. , & Liang, J. . (2017). Fine-Tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[8] Rajpura, P. , Aggarwal, A. , Goyal, M. , Gupta, S. , Talukdar, J. , & Bojinov, H. , et al. (2017). Transfer learning by finetuning pretrained cnns entirely with synthetic images.
[9] Galdi, C. , Nappi, M. , & Dugelay, J. L. . (2017). Secure User Authentication on Smartphones via Sensor and Face Recognition on Short Video Clips.International Conference on Green. Springer, Cham.
[10] Xusong Chen, Dong Liu, Zhengjun Zha &Wengang Zhou (2018), Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction, the 2018 MM
[11] Vila, Marius, Bardera, Anton, Feixas, & Miquel, et al. (2013). Tsallis entropy-based information measures for shot boundary detection; and keyframe selection. Signal Image & Video Processing,7(3), 507-520.
[12] Hannane, R. , Elboushaki, A. , Afdel, K. , Naghabhushan, P. , & Javed, M. . (2016). An efficient method for video shot boundary detection and keyframe extraction using sift-point distribution histogram. International Journal of Multimedia Information Retrieval, 5(2), 89-104.
[13] Zhuang, Y., Yong, R., Huang, T. S., & Mehrotra, S. (2002). Adaptive key frame extraction using unsupervised clustering. International Conference on Image Processing.
[14] Li, C. , Wang, P. , Wang, S. , Hou, Y. , & Li, W. . (2017). Skeleton-based action recognition using lstm and cnn.
[15] Wang, X. , Gao, L. , Song, J. , & Shen, H. . (2016). Beyond frame-level cnn: saliency-aware 3d cnn with lstm for video action recognition. IEEE Signal Processing Letters , PP(99), 1-1.
[16] Tran, D. , Bourdev, L. , Fergus, R. , Torresani, L. , & Paluri, M. . (2014). Learning spatiotemporal features with 3d convolutional networks.
[17] Saleem, Summra; Hassan, Muhammad A.; Khan,2018, Learning deep C3D features for soccer video event detection, ICET 2018.

[18] Ji, S. , Xu, W. , Yang, M. , & Yu, K. . (2013). 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence,35(1), 221-231.
[19] Simonyan, K. , & Zisserman, A. . (2014). Two-stream convolutional networks for action recognition in videos.
[20] Dai, J. , Qi, H. , Xiong, Y. , Li, Y. , & Wei, Y. . (2017). Deformable Convolutional Networks. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE.