Temporal-scale Convolutional Networks for Human Action Recognition Based on Key-Frame Extraction

Zhao-qiang WEI, Yong-qiang KONG, Zhen-gang WEI* and Xiao-long ZHANG

College of Information Science and Engineering, Ocean University of China, Qingdao, China

*Corresponding author

Keywords: Convolutional networks, Action recognition, Key-frame, Video segment.

Abstract. Human action recognition is an important part of intelligent video analysis. Recently, deep learning has made significant progress in this field, and state-of-the-art methods are based on the two-stream convolutional networks. In long-term action recognition, existing approaches mainly use video frames obtained by averaging or sampling as input, which may lose important information in the sampling interval. By defining the amount of video information, we propose a method of segment division and key frame extraction for action recognition is proposed, where multi-temporal-scale two-stream networks are used to extract features. We achieve 94.2% accuracy on a widely used action recognition benchmark (Split1 of UCF-101).

Introduction

Human action recognition is an important part of intelligent video analysis. The core is to segment video into image sequence and classify the sequence by extracting the temporal and spatial features. In the early years, hand-crafted features were commonly used to achieve action classification, such as using spatio-temporal descriptors [1]. In recent years, with the great success of the deep neural network in the field of image recognition, the use of deep learning for video behavior recognition has achieved remarkable results.

One of the major difference between video and static images is that video has temporal components. To extract temporal features, video action recognition needs to deal with a large number of video frames and get implied cues. Existing researches usually divide the video into several segments, and extract features of each detected segment. The early methods often use a single Convolutional Network (ConvNet) to extract features, and the results are obtained based on the mean value of the extracted results. The two-stream method [2] introduce optical flow as an auxiliary input, and use two ConvNet architectures to extract the temporal as well as the spatial information in the segments, respectively. LRCN [3] introduced a Recurrent Neural Network (RNN) to extract time characteristics. The above methods use average sampling to obtain the frame which is need to be detected. Most existing researches focus on how to extract temporal information, with little attention to the quality of detected frames. Whether the detected frame contains the key information of the current action is an important factor that affects the result of action recognition. Here we call the video frame or optical flow sequence that contains key action information “key frame”, and call the sampling interval “segment”.

The difference between video and static images is not just the temporal components. Because video comes from continuous shooting, it is easy to have disturbances such as motion blur, loss of focal length and so on. Even video compression can bring noise. However, there is a difference in the amount of information between different video frames. Meanwhile, the amount of motion information contains among different video segments is often very different because of the uneven movement of natural behavior. Traditional key frame extraction methods are mostly based on global feature [4], such as brightness histogram and gradient histogram. It is difficult to find fragments of key information in action sequence with small scene change.
In order to solve the above problems, we propose a video segmentation method based on the amount of motion information and a key frame extraction method based on the amount of image information. On the basis of both of the two cues, we design an action recognition framework using multi-scale temporal network based on the two-stream approach. The proposed method performs well on the UCF101 dataset, which achieves a high recognition accuracy rate of 94.2%.

### Segment Split

Motion information is often uneven in temporal domain. When the action is detected, the main body will occupy most of the space of the effective video, and the dense optical flow can express the motion state of the main body. The definition of dense optical flow is as follows:

\[
T(x, y) = I(x + u, y + v)
\]  

where T and I represent the past and the current frames respectively. u and v distribution represent the offset of pixels on x and y axis. The optical flow image is the matrix of the coordinates x, y, which corresponds to u and the v. When the main motion is obvious, the optical flow image should have a higher absolute value. Meanwhile, if the main body moves significantly, it will contain more behavioral information.

The distribution of action information in videos is uneven. To make the motion information more uniform, the variance DM of the segments needs to be minimized when the number of segments NS is fixed, which is defined as follows:

\[
D_M = \sum_i (M_i - M') / N_s
\]  

where Mi represents the amount of motion information of the i-th segment, M’ is the average information of the segment. Information content M of each fragment is expressed as follows:

\[
M = \sum_i \sqrt{\sum_c \sum_{x,y} FL_i^c(x,y,c)}
\]  

where FL is the 2-channel optical flow image of the segment, and c represents flow channel.

Exact solution can be obtained by using dynamic programming. However, optical flow often introduces errors due to the fluctuation of video quality. For computational efficiency, we use greedy algorithm to obtain approximate solutions.

Figure 1. Examples of edge extraction.
Key Frame Extraction

The pre-trained deep ConvNet model can greatly improve the performance of the classifier. It is generally believed that this is because the networks learn representations of large image datasets. The samples of large datasets are clearer, the learning of fuzzy features is therefore less. Blurred images may affect the final recognition results. In addition, images which contain more objects usually contain more information. In order to solve these problems, we use an image information evaluation technique to find the frames that contain the largest information in video segments.

Many researches show that the edge of the image is more concerned by human vision. There are several state-of-the-art object detection methods based on image edge extraction [3]. Considering that the average edge intensity can effectively represent the amount of frame information. We define the amount of information $E_i$ for the $i$-th frame:

$$E_i = \frac{\sum_{x,y} ED_i(x, y)}{W \times H}$$  \hspace{1cm} (4)

where $ED$ is the gray edges obtained from frames, and $W$ and $H$ are width and height respectively. We use the edge extraction technique proposed in [6] to obtain edges. This method runs faster and the gray level of the image can smoothly express the blur of the object. Examples of edges extraction are shown in Figure 1. The frame with the most information in the segment is taken as the key frame:

$$idx = \arg \max(E)$$  \hspace{1cm} (5)

Action Recognition Framework

Feature Extraction

The two-stream method [2] uses image classification network to construct spatial and temporal networks, respectively. This architecture performs well on human action recognition. Moreover, its memory occupancy size and computing scale are comparable with the commonly used image classification network. The two-stream method can also use the model pre-trained on large dataset (such as ImageNet), which not only greatly improves the classification accuracy, but also prevents overfitting. We use this spatio-temporal model to extract video features and use key frames of each segment in the video as input.

Spatio-temporal Model

Spatial Net. The spatial part is the image classification network. We choose BN-Inception [7] which uses convolution kernel of different scales and has strong multi-scale detail recognition ability. The use of Batch Normalization (BN) makes this model converge more rapidly and accurately. The fixed size single RGB images are inputs of the network. We use a model pre-trained on Pascal VOC 2012 dataset [8] to initialize the convolution parameters.

Temporal Net. The temporal part is the optical flow network. We use the modified BN-Inception architecture to implement it. Based on the original network, we expand weights of the first convolutional layer, with the aim of making the inputs to support more channels. Specifically, we first aggregate the parameters of the first convolution layer across channel. After that, the new parameters divided by the number of the channels. The results are copied across channel as the new parameters of this layer. The input channel is fixed to 10, which is Obtained by stacking 5 dense optical flow over xy direction.

Since the existing optical flow networks only extract information from continuous frames, we introduce a multi-temporal-scale optical flow network to extract features from different temporal scales. By uniformly sampling at different time span (i.e., $FL_{t-k^*}, t \in \{0, \ldots, 4\}$), we obtain multiple
optical flow sequences (with the same length) at the same end time. \( t \) and \( k \) are the current time and sampling interval, respectively. Dense optical flow is stacked as input to the temporal network. During testing, multi-scale flow sequence is tested separately, and the classification result is obtained based on average pooling.

![Figure 2. Training of the two-stream network.](image)

**Features Fusion.** The training of the spatio-temporal network is shown in Figure 2. The selected video frames in each segment are used as input, and the outputs of the last convolutional layer are used as outputs of the classifier. For networks fusion, we use the average pooling approach to achieve the goal.

During training, video segments are divided based on the amount of motion information. Frames are randomly selected as inputs. We perform average pooling to the spatiotemporal features of each segment (i.e., the convolution map). The pooling results are then used as input of the loss function allow backpropagation. The loss function in our proposed spatiotemporal architecture is based on softmax cross entropy. Training is implemented using Stochastic gradient descent. During testing, we divide more segments; key frames are obtained according to the amount of image information. Then these frames are processed by the classifier. The final predictions are obtained by the average of all results with softmax.

**Experiments**

UCF101 [9] is a widely used human action recognition benchmark, which contains 101 categories of 13320 videos. The most commonly used training/test set is Split1.

**Implementation Details**

Several works proved that segmented fusion training can effectively improve the representation ability of the action model [10]. However, it is not that the more segments corresponds to better performance. During backpropagation, intermediate data need to be retained. Therefore, the number of segments is inversely proportional to the batch size. Batch too little can lead to slow convergence. We thus need to find a balance between the number of segments and batch size. The number of segments used in our experiment training is set to 3, and the batch size is set to 32. We used the Caffe toolbox [12] to conduct all experiments.

![Table 1. Accuracy on UCF101, Split1.](table)

<table>
<thead>
<tr>
<th>Test setup</th>
<th>Frame</th>
<th>3 segments</th>
<th>10 segments</th>
<th>25 segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average segments</td>
<td>First</td>
<td>92.3%</td>
<td>93.8%</td>
<td>93.6%</td>
</tr>
<tr>
<td>Average segments</td>
<td>Key</td>
<td>92.8%</td>
<td>94.2%</td>
<td>94.0%</td>
</tr>
<tr>
<td>Information segments</td>
<td>Key</td>
<td>92.8%</td>
<td>94.4%</td>
<td>94.1%</td>
</tr>
</tbody>
</table>
Test Results
Table 1 shows the test results in terms of recognition accuracy, where the number of segments are 3, 10 and 25 respectively. The presented results are obtained from fused two-stream networks. And the blank frame is removed.

The action classifier achieves the best performance when the number of segments is set to 10. In the case of different number of segments, both the information based segmentation and key frame selection methods can improve the performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>Year</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stream[2]</td>
<td>2014</td>
<td>88.0%</td>
</tr>
<tr>
<td>Two-stream fusion[11]</td>
<td>2016</td>
<td>92.5%</td>
</tr>
<tr>
<td>TSN [10]</td>
<td>2016</td>
<td>93.5%</td>
</tr>
<tr>
<td>Temporal Inception [5]</td>
<td>2017</td>
<td>94.2%</td>
</tr>
<tr>
<td>Ours</td>
<td>2018</td>
<td>94.4%</td>
</tr>
</tbody>
</table>

Comparison with the State-of-the-art
We compare the proposed method with 4 state-of-the-art action recognition models. Results are shown in Table 2. [2] is the first two-stream ConvNets, which is comprised of RGB and optical flow networks. Two-stream fusion [11], TSN [10] and Temporal Inception [5] are all based on the two-stream architecture. After introducing the segment splitting and key frames extraction methods, our proposed action classifier achieved the highest recognition accuracy of 94.4%, which is better than other methods. Our method also has other advantages. For example, Temporal Inception fuses features by training a temporal domain Inception network which is performed between segments. Our method achieves the higher accuracy rate only use average fusion, which indicates that there is still room for improvement.

Conclusion
By analyzing the current video behavior recognition method, we propose a video fragment partition method based on the amount of motion information and a key frame extraction method based on the amount of image information. On the basis of both of them, we design a behavior recognition system using multi-scale optical flow network combined with dual flow network. The experiments on the UCF101 dataset demonstrate that the segmentation and key frame selection methods proposed by us can effectively improve the accuracy of the behavior recognition and the best results on the data slice.

Acknowledgement
This research was supported by the National Natural Science Foundation of China (nos. 60374031).

References


