Human Activity Recognition Based on Two Stream Convolutional Neural Network

Ning Zhang*, Xiaochen Li**, Suk-Hwan Lee***, Tae-Kook Kim*, Eung-Joo Lee*
*Dept of Information Communication Engineering, Tongmyong University, Korea
**Dept of Mechanical Engineering, Dalian Polytechnic University, China
***Dept of Information Protection Engineering, Tongmyong University, Korea

Abstract
In this paper, we focus on the action classification problem in video. Here we use two independent CNNs to separate the spatial information and time information spatial and temporal in the video, and then we integrate late fusion. The spatial stream recognizes motion from each frame of the video, and the temporal stream recognizes the motion by inputting motion information of the dense optical flow. Both streams are done through the CNN network. By separating the time and space information, you can use the off-the-shelf database to train the two networks.

1. Introduction

Video understanding is currently a very hot and challenging direction in the field of computer vision. The video understanding direction includes many sub-studies. Taking the ACTIVITYNET organized by CVPR as an example, a total of five Tasks were proposed in 2017. Task 1: Untrimmed Video Classification; Task 2: Trimmed Action Recognition; Task3: Temporal Action Proposal; Task4: Temporal Action Localization; Task5: Dense-Captioning Events. The research report focuses on behavioral recognition and behavioral testing. That is, Task 2 and Task 4 in the above task description.

Trim하ed Action Recognition is an important issue in the direction of video understanding and has been studied for many years. After deep learning, the problem was gradually solved, and now it has achieved satisfactory results in the data set. The behavior classification problem is simply: for a given segmented video segment, it is classified according to its human behavior. For example, girl makeup, boys playing, running, etc. This task does not need to determine the start time and end time of the behavior in the video.

Before the advent of deep learning, the best performing algorithm was iDT[1][2], and the subsequent work was basically improved on the iDT method. IDT's idea is to use the optical flow field to obtain some trajectories in the video sequence, and then extract the HOF, HOG, MBH, and trajectory four features along the trajectory. The HOF is based on the gray-scale image calculation, and the others are based on the dense optical flow. Finally, the features are encoded using the Fisher Vector method, and the SVM classifier is trained based on the coding training results. After deep learning, there are many ways to solve this problem, including: Two-Stream [3][4], Convolution 3 Dimension [5], and RNN [6] direction.

2. Traditional Methods

The iDT (improved Dense Trajectories) method is the most classic method before deep learning. Although the current method based on deep learning has surpassed iDT, the idea of iDT is still worth learning, and there are always some improvements after doing ensemble with the results of iDT. The idea of iDT is mainly reflected in the two articles "Dense Trajectories and Motion Boundary Descriptors for Action Recognition" and "Action Recognition with Improved Trajectories".

2.1 Densely Sampled Feature Points

As shown in Figure 1 below, the iDT algorithm framework mainly includes three parts: dense sampling feature points, feature trajectory tracking and trajectory-based feature extraction.

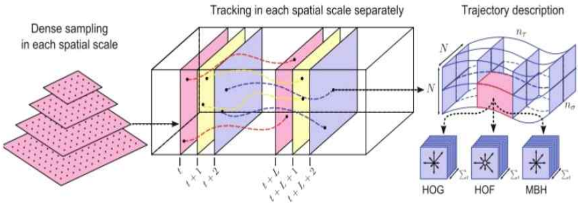

Figure 1. iDT algorithm architecture diagram

The DT method densely samples feature points in a multi-scale image by meshing.

2.2 Trajectory and Track Descriptor

Assume that the coordinates of a feature point densely sampled in the previous step are $P_t = (x_t, y_t)$, and then use the following formula to calculate the position of the feature point in the next frame image.

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M \times w_t)|_{x_t, y_t} \qquad (1)$$

In the above formula, w_t is a dense optical flow field, which is calculated by I_t and I_{t+1}. M stands for the median filter and has a size of 3x3, so this formula obtains the motion direction of the feature point by calculating the median value of the optical flow in the feature point field.

2.3 Motion Descriptor

In addition to the trajectory shape feature, more powerful features are needed to describe the optical flow. The HOF, HOG and MBH features are used in DT/iDT. The following is a brief description of these features.

HOG feature: The HOG feature calculates a histogram of the gray-scale image gradient. The number of bins in the histogram is 8. Therefore, the length of the HOG feature is 2*2*3*8=96.

HOF feature: HOF calculates the histogram of the optical flow. The number of bins in the histogram is taken as 8+1, and the first 8 bins are the same as the HOG. An additional one is used to count pixels whose optical flow amplitude is less than a certain threshold. Therefore, the feature length of HOF is 2*2*3*9=108.

MBH feature: MBH calculates the histogram of the optical flow image gradient, which can also be understood as the HOG feature calculated on the optical flow image. Since the optical flow image includes the X direction and the Y direction, MBHx and MBHy are calculated separately. The total feature length of MBH is 2*96=192.

Finally, the normalization of the features is performed. The DT algorithm normalizes the HOG, HOF and MBH using the L2 norm.

3. Two Stream Convolutional Neural Network

The Two-Stream method is a major trend in deep learning in this direction. It was first proposed by the VGG team on NIPS [3]. In fact, some people have tried to use deep learning to deal with behavior recognition before. For example, Li Feifei's team [7], through superimposed video multi-frame input to the network for learning, but unfortunately this method is worse than manually extracting features. When Two-Stream CNN comes out, it means that deep learning has taken a big step in behavior recognition.

The Two-Stream CNN network is divided into two parts as the name suggests, one part processing RGB images and the other part processing optical flow images. Ultimately joint training and classification. This article has three main contributions.

First, the paper proposes a two-stream CNN network, which consists of two dimensions: space (RGB) and time (optical flow).

Secondly, the authors propose to use the network to train multi-frame density optical flow, which can be used as input to achieve good results with limited training data.

Finally, the multi-task training method is used to combine the data sets of the two behavioral classifications, increase the training data, and finally achieve better results on both data sets.

Because video can be divided into two parts: space and time. In the space part, each frame represents spatial information, such as targets, scenes, and so on. The time part refers to the motion between frames, including the motion of the camera or the motion information of the target object. Therefore, the network consists of two parts, which deal with the two dimensions of time and space.

Each network consists of CNN and the final softmax. The final softmax fusion mainly considers two methods: averaging, training an SVM on the stacked softmax. The structure of the network is shown below.

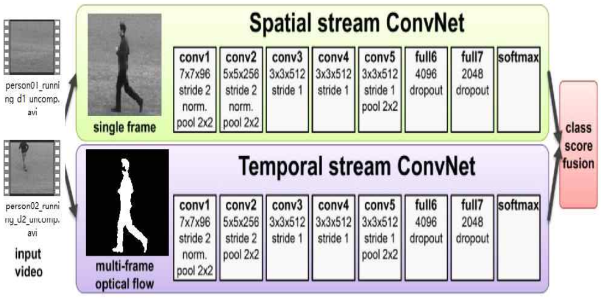

Figure 2. Two-stream network structure diagram

Optical flow stacking, or a simple superposition of optical flows. Simply put, the optical flow between every two frames is calculated, and then simple stacking.

Suppose we consider the classification of actions (behavior recognition mainly consists of two directions, one is action classification, giving a video truncation, judging the action category of the video, or offline. The other is motion recognition, giving a natural video, no For any cropping, you need to know the start time and end time of the action, and then know the type of action. Consider encoding a small segment of video, assuming a starting frame of T and a continuous L frame (without T frames). Calculating the optical flow between two frames, you can finally get L optical flow fields, each optical flow field is 2 channels (because each pixel has a movement in the x and y directions).

$$I_\gamma(u,v,2k-1) = d^{x_{\gamma+k-1}}(u,v) \qquad (2)$$

$$I_\gamma(u,v,2k) = d^{y_{\gamma+k-1}}(u,v), \\ u=[1;w], v=[1;h], k=[1;L] \qquad (3)$$

Finally, we input these optical flow fields to get the corresponding feature maps.

5. Experimental Results

In the end, the method achieved the best consistent effect with the iDT series on UCF-101 and HMDB-51. The accuracy on UCF-101 is 88.0%, and the accuracy on HMDB is 59.4%.

Table 1. Two-stream experiment result

Method	UCF101	HMDB51
Improved dense trajectories	85.9%	57.2%
iDT with higher-dimensional encodings	87.9%	61.1%
iDT with stacked Fisher encoding	-	66.8%
Spatio-temporal HMAX network	-	22.8%
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model(averaging)	86.9%	58.0%
Two-stream model(SVM)	88.0%	59.4%

Acknowledgment

본 연구는 과학기술정통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음"(2019-09). 그리고 이 성과는 2019년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2017R1C1B2011285).

References

[1] Wang H. and Schmid C., "Action recognition

with improved trajectories," *2013 IEEE International Conference on Computer Vision (ICCV)*, pp. 3551-3558.

[2] Wang H, Kläser A, Schmid C, et al. "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, Vol. 103 No. 1, pp. 60-79, 2013

[3] Simonyan K and Zisserman A. "Two-stream convolutional networks for action recognition in videos," *2014 Advances in Neural Information Processing Systems*. pp. 568-576.

[4] Feichtenhofer C, Pinz A, Zisserman A. P. "Convolutional two-stream network fusion for video action recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1933-1941

[5] Tran D, Bourdev L, Fergus R, et al. "Learning spatiotemporal features with 3d convolutional networks," *2015 IEEE International Conference on Computer Vision (ICCV)*. pp. 4489-4497.

[6] Du W, Wang Y, Qiao Y. "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3725-3734.

[7] Karpathy A, Toderici G, Shetty S, et al. "Large-scale video classification with convolutional neural networks," *2014 IEEE conference on Computer Vision and Pattern Recognition*. pp. 1725-1732.