3D 컨벌루션 신경회망의 특강 융합 기반의 인간 행동 인식

오정정, 김태국, 이응주

동명대학교 정보통신공학과

e-mail : ejlee@tu.ac.kr

Human Action Recognition Based on Feature Fusion of 3D Convolution Neural Network

Tingting Wu, Tae-Kook Kim, Eung-Joo Lee

Dept. of Information Communication Engineering, Tongmyong University

Abstract

In the existing 3D convolution neural network, when convoluted in the time dimension, the 3D convolution kernel is used to convolute successive frame image cubes in the form of sliding windows. However, due to the changes of pedestrian posture, action and position, the convolution at the same place is inappropriate, and when the 3D convolution kernel is convoluted in time dimension, only time dimension features of three consecutive frames can be extracted in time dimension, which is not a good enough to get action information. Therefore, in order to increase action information, enrich action features and enhance the robustness of single feature representation, this paper proposes an action recognition method based on feature fusion of 3D convolution neural network.Based on the VGG16 network model, sending a per-acquired optical flow image for learning,then get the time domain features, and then the feature of the time domestication is extracted from the features extracted by the 3D convolution neural network. Finally, the behavior classification is done by the SVM classifier.

1. Introduction

In computer vision, human action recognition involves pattern recognition, image processing, computer vision, artificial intelligence and other fields.

It is widely used in human-computer interaction, action capture analysis, video monitoring and safety, environmental control detection and prediction. At present, human action recognition is mainly affected by individual differences, angle of view changes, camera movement and illumination angle, it is still a challenging subject to accurately identify and analyze human behavior in real scenes, so it is very important to develop a set of advanced action recognition algorithm.

With the emergence of a large number of video data sets for action recognition, the research on How to extract effective features from the video is the key to solve the above problems and design more effective behavior recognition framework.

At present, methods of human behavior recognition are mainly divided into the methods based on traditional behavior recognition and the methods based on deep learning. Recognition has encountered serious computational burden. How to extract effective features from video is the key to solve the above problems and design more effective recognition framework.

At present, methods of human behavior recognition are mainly divided into the methods based on traditional behavior recognition and the methods based on deep learning.

The traditional action recognition method is mainly composed of two steps. The first step is to extract the features of video image; the second step is to use the learning classifier to classify the features. In the real scene, different behaviors have obvious differences in appearance and movement mode. So it is difficult to choose the appropriate features, and deep learning model can learn by sample characteristics, which have the advantage of better than traditional action recognition method.In recent years, with the great success of deep learning method in image classification and target detection, people begin to use deep learning method in video behavior recognition.

2. Time Domain Feature Extraction

In order to extract better time domain features, the model design of convolution neural network is particularly critical. This paper is based on VGG16 network model, which was published in 2014. The network shows that stacking multiple layers is a key factor in improving computer vision performance.

The network features smaller size of convolution kernel, smaller design of convolution step, smaller lower sampling window and deeper network structure. It is mainly composed of small 3×3 convolution operation and 2×2 pooling operation. In order to preserve the spatial resolution, edge processing is carried out in each convolution layer, that is, the size of the image is not changed during the convolution processing.

The advantage of VGG network is that stacking multiple small convolution kernels without using pooling operation can increase the representation depth of the network and limit the number of parameters at the same time. First, it combines three non-linear functions to make the decision function more perceptive and representational. Second, the parameters were reduced by 81 percent, while the receptive field remained unchanged.In addition, the effectiveness of different convolution kernels is improved.

3. Feature Fusion

The fusion method is mainly divided into two aspects: feature fusion and result fusion.The methods of feature fusion are serial feature fusion, weighted feature fusion, and serial feature fusion or weighted feature fusion based on a series of feature correlation coefficients derived from the two fusion methods.Serial feature fusion method enhances the robustness of single feature representation and achieves good recognition effect in the field of behavior recognition. At the same time, it has the advantages of simple fusion and simple calculation. However, serial feature fusion will increase feature dimension, which may lead to large error in learning results.

In this experiment, we successively adopted direct serial feature fusion and direct weighted feature fusion.

(1) serial feature fusion the method of serial feature fusion is to combine two sets of features in the sample space directly into a new feature vector, and then extract and compress the synthesized feature vector.

Suppose $trainX$ and $trainY$ are two different groups of characteristics, and the feature size of $trainX$ is $M_1 \times N_1$, represented by $trainX_{M_1 \times N_1}$; and the feature size of $trainY$ is $M_2 \times N_2$, represented by $trainY_{M_2 \times N_2}$. The combined feature $fusion_{concat} = [trainX, trainY]$ and requirements $M_1 = M_2 = M$. The combined feature size is $M \times (N_1 + N_2)$. In the experiment, the obtained 3D convolution feature size is 147926×4096 and the time-domain feature size is 147926×4096 , so the feature size after direct serial feature fusion is 147926×8192 .

(2) Weighted feature fusion weighted feature fusion method is to set different weights for the two groups of features in the sample space according to the proportion of feature participation, and then merge the features.

Suppose $trainX$ and $trainY$ are two different groups of characteristics, and the feature size of $trainX$ is $M_1 \times N_1$, represented by $trainX_{M_1 \times N_1}$; And the feature size of $trainY$ is $M_2 \times N_2$, represented by $trainY_{M_2 \times N_2}$. The weights of the two groups of features are w_1, w_2. Requirements $M_1 = M_2 = M$, $N_1 = N_2 = N$, and the combined feature

$fusion_{\text{weight}} = [w_1 \times trainX + w_2 \times trainY]$. The combined feature size is $M \times N$. In the experiment, the obtained 3D convolution feature size is 147926×4096 and the time-domain feature size is 147926×4096 , so directly weighted feature fusion feature size is 147926×4096 .

4. Experiment Analysis

In this paper, SVM is adopted to classify the fusion features generated by each model, the UCF101 data set contains 101 types of behaviors. Each type of behavior was performed by 25 different groups of people, and everyone has multiple video. Taking the first seven groups of each type of behavior as test samples, the last 18 groups were used as another test sample.

The test data is sent into the trained SVM classifier, each classifier classifies and identifies the test data, it is considered to belong to the category that is classified into the most times. Finally, compared with the labels marked in advance, if the categories are consistent, the classification is considered to be correct.

From the experimental results, the accuracy of feature fusion is greatly improved compared with 3D convolution network. The weight setting strategy has the highest accuracy of weighted feature fusion. The serial feature fusion has a slightly higher accuracy than the weighted feature fusion of the first weight setting strategy. However, because the feature dimension is doubled after serialization, the feature size is increased, so it takes more time to run.

References

[1]Kanade T. A System for Video Surveillance and Monitoring. Proceedings of Vsam Final Report Carnegie MellonUniversity Technical Report. 2000, 59(5)：329-337.

[2]Bird N, Atev S, Caramelli N, et al. Real time. Proceedings of online detection of abandoned objects in public areas. 2006：3775-3780.

[3]Laptev I,Marszalek M, Schmid C, et al. Learning realistic human actions from movies. Proceedings of IEEE Computer Vision and Pattern Recognition, 2008. CVPR 2008. 2008: 1-8.

[4]Kovashka A, Parikh D, Grauman K. WhittleSearch, Interactive Image Search with Relative Attribute Feedback. Proceedings of Kluwer Academic Publishers, 2015.

[5]Matikainen P, Hebert M, Sukthankar R. Trajectons. Action recognition through the motion analysis oftracked features[C]. Proceedings of IEEE International Conference on Computer Vision Workshops. 2009: 514-521.