Action recognition based on element-level fine-grained multi-modal fusion

Guozheng Peng\textsuperscript{1∗}, Lixin Han\textsuperscript{1} and Jiaxue Yang\textsuperscript{1}

\textsuperscript{1}School of Computer and Information, Hohai University, Nanjing, Jiangsu, 211100, China
\textsuperscript{*}Corresponding author’s e-mail: jerrypeng@hhu.edu.cn

Abstract. Traditional action recognition algorithms often only pay attention to video RGB features or optical flow features. These methods do not make good use of the audio information in the video. Based on RGB and optical flow characteristics, this paper introduces the processing of audio information, and classifies videos based on element-level fine-grained multi-modal fusion. Through experimental comparison, the accuracy of the multi-modal fusion algorithm proposed in this paper is improved by 7.38\% on the HMDB51 dataset and 3.18\% on the UCF101 dataset compared to the simple modal splicing. At the same time, it is proved that the introduction of audio modes can effectively improve the performance of the model.

1. Introduction
Action recognition is to recognize human actions from a video. However, the video gathers rich content information such as visual information and auditory information. Research shows that different modalities can affect the accuracy of action video recognition, so how to be effective Extracting content features and integrating them becomes the research focus. In recent years, a large number of scholars have paid attention to the research based on multi-modal action recognition and have produced many excellent results, but there are still some shortcomings. Existing action recognition often uses the visual information of the video when analyzing the features of the video content, and treats the extraction of the visual features of the video as image feature extraction for processing, and lacks the use of the rich information contained in the video frame. In addition, the optical flow information of the video can reflect the motion information in the video, and the audio branch can contain the emotional representation of the video. The characteristics of the two branches can bring positive benefits to the video understanding, so they cannot be ignored. In the feature fusion stage, the existing fusion methods are difficult to analyze the relationship between modal branches.

In response to the above problems, this article proposes some improvement methods.

(1) Aiming at the problem of insufficient utilization of video frame-level information, this paper first uses the inter-frame difference method to extract the key frames corresponding to the video, uses the FlowNet network to extract the optical flow information corresponding to the video, and processes the RGB spatial information and optical information of the video key frames in parallel. Stream information.

(2) For each key frame, this paper introduces the AdaScan algorithm \cite{1} to analyze its contribution to the understanding of video semantics, so that the model tends to strengthen the frames that have a greater impact on the understanding of video semantics while ignoring redundant frames.
(3) In view of the existing fusion methods that are difficult to analyze the related information between modal branches, this paper proposes an element-level fine-grained feature fusion method, calculates the element-level influence of each modal branch, and fuses them with weight.

2. Feature extraction and fusion

2.1. Visual feature extraction

2.1.1 Key frame selection
Video in real life contains a very large number of frame sequences, and there are very serious redundancy problems between consecutive video frames. In order to reduce the computational complexity, it is necessary to filter key frames. Because the video frame sequence has continuous characteristics and can express the characteristics of the video in the time sequence dimension, the key frame selection should be able to reflect the timing dependence of the relevant frame on the video on the side, and the sampling should save as much as possible the summary of the video content of the key frame sequence ability to avoid the loss of main information. The existing key frame selection algorithms include uniform sampling, continuous sampling, frame difference method and clustering-based methods. Continuous sampling usually selects the starting position of sampling in a random manner. From this position, several consecutive video frames are taken. It can be found that continuous sampling cannot fully express the video content, and serious information loss may occur. Uniform sampling is to extract frames from the video at a certain interval. Extracting frames without filtering is prone to individual frame redundancy and some key information loss. The clustering method treats all the frames in the video as a set. By calculating the similarity between the frames, the clustering algorithm is used, and finally the center of the cluster is selected as the key frame. It can be found that a group represents a person lifting up in the video The hand movements are all regarded as similar frames, and the clustering algorithm often chooses one frame to extract, which cannot solve the timing problem in the video. Considering comprehensively, this paper chooses the inter-frame difference method as the method of extracting video key frames. First, read the video file, make the difference between the pixels at the corresponding positions of two adjacent frames, as the difference strength of the two frames, and then select the video frame with the inter-frame difference greater than the threshold as the key frame. However, it should be noted that this article does not address short videos such as gesture recognition and action classification. In long videos, there are often scene changes. At this time, the differential strength may be too large. In order to make the sampling method suitable for longer videos, it is understood that this article also adds the upper threshold of the threshold to filter the sudden changes between frames (such as scene changes, etc.).

2.1.2 Optical flow information
In this paper, the FlowNet network is selected as the optical flow extraction algorithm. The network is divided into the encoding module shown in Figure 1 and the decoding module shown in Figure 2. The encoding module part includes 9 convolutional layers and a ReLU function activation layer, and the decoding part includes 4 deconvolutions and a ReLU function activation layer. The network directly stacks two frames of images as the input of the encoding module; and in the decoding module, for each deconvolution layer, the output of the previous layer is input at the same time, the optical flow predicted by the network for the previous layer, and the encoding module The feature layer, in this way, obtains deeper abstract information and shallow specific information of the image to make up for the information lost due to the reduction of the feature space scale.
2.1.3 Feature extraction

The spatial features of video mainly include scene, texture and edge features, which can generally be well understood by processing RGB images. Therefore, this paper uses the key frames extracted by the above steps to extract video spatial branching features. Literature [2] pointed out that usually only a part of the frames in a video contain useful information for video understanding. At the same time, the AdaScan network is proposed to extract the spatial characteristics of the video by traversing the video frames and overlaying their relative importance with weights. The idea of the AdaScan algorithm is to determine the importance of each frame of video after receiving the characteristics of each frame of the video. For video frames with minimal recognition effect, the author assigns them a very small weight to achieve the goal of "discarding" useless frames effect. The key module of the algorithm is adaptive pooling, which recursively predicts the importance of the recognition and understanding of the current frame based on the characteristics of the current frame and the frame set vector up to the current frame. In this way, the video frames that have a greater influence on the recognition video are gathered, reducing the overall impact of redundant frames and reducing noise.

The AdaScan network structure is shown in Figure 3. This article divides it into two branches, which are used to process RGB frames and optical flow information. The following uses RGB frames as an example to illustrate the calculation process of the AdaScan algorithm. The optical flow information is similar.
2.2. Audio feature extraction
Audio data cannot be directly input to the network for calculation, so the audio data needs to be preprocessed before use. This article first resamples the audio file, then uses the Fourier transform to obtain the corresponding spectrogram, and finally uses the VGGish network to extract the characteristics of the spectrogram as the audio characteristics. After the spectrogram corresponding to the audio file is obtained, it is input into the VGGish network.

2.3. Feature fusion
The existing multi-modal fusion methods are mainly divided into early fusion and late fusion. Early fusion [3-4] extracted the feature vector expressions of each modal branch, and spliced them as the input of the final classifier, but the early fusion ignored the connection between the modal branches, and had some expression ability lack; late fusion [5-7] trains the model separately on each modal branch and obtains the corresponding results, and finally takes the weighted average or takes the maximum value as the output result of the model. However, neither early fusion nor late fusion can analyze the correlation between modal branches at a fine-grained level. The relationship between the various modal branches of the video is complicated, and the understanding of the semantics of the video content is not the same. This article proposes an element-level feature fusion method as shown in Figure 4 to analyze the modal branches at a fine-grained level. Analyze the influence of element level between different modal branches in parallel, strengthen the influence of main elements on the model, and suppress noise data at the same time.
3. Experiment and result analysis

3.1. Dataset

This article uses a total of two data sets, namely HMDB51 [8], UCF101 [9]. The video data in UCF101 is collected from YouTube. The data set has 13,320 videos in total, which are divided into 101 categories. UCF101 is currently one of the video classification data sets with the largest number of action categories and samples. The 101 action category videos in this data set are divided into 25 groups, and each group roughly contains 4-7 action categories.

Most of HMDB51's data are clips from movies, and some are from public databases, such as Prelinger archives, YouTube videos, and Google videos. The data set contains a total of 6849 video clips. These video clips are divided into 51 action categories, and each action category contains at least 101 video clips.

3.2. Experiment results

This paper compares the two data sets of HMDB51 and UCF101, and combines the feature fusion algorithm proposed in this paper with RGB, RGB and Flow information, as well as simple mosaic of RGB, Flow and Audio respectively. By comparing the experimental results of RGB, RGB+Flow and RGB+Flow+Audio, it is found that the increase in the number of modes can effectively improve the accuracy of the results, and the increased modes make up for the information not contained in other modes to a certain extent. Comparing the experimental results of the proposed algorithm and RGB+Flow+Audio, it is found that the element-level feature fusion algorithm proposed in this paper can effectively improve the accuracy of the model.

Table 1. Classification accuracy rate using different features under two kinds of data sets (%)

| Dataset    | RGB  | RGB+Flow | RGB+Flow+Audio | Ours |
|------------|------|----------|----------------|------|
| HMDB51     | 31.34| 41.44    | 54.92          | 62.30|
| UCF101     | 66.19| 73.20    | 82.56          | 85.74|

4. Conclusion and prospect

Traditional action recognition algorithms have problems such as insufficient use of video frame-level features or excessive redundant information, and insufficient fusion of visual information and auditory information. Therefore, this chapter first uses video key frames to extract visual features, and then uses AdaScan to perform. The secondary screening is mainly divided into two aspects: RGB spatial
information and optical flow information, and then fully integrated with the auditory information extracted through the pre-training network.

In the feature fusion stage, this paper proposes an element-level feature fusion method to analyze the relationship between modal branches, analyze the influence of different modal branches at the element level in parallel, and add entropy regularization to help the model tend to choose Elements with greater impact on video recognition are discarded, and elements with less impact are discarded to achieve the purpose of element screening.

References
[1] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE transactions on knowledge and data engineering, 2005, 17(6): 734-749.
[2] Kar A, Rai N, Sikka K, et al. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos[C].Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, 2017: 3376-3385.
[3] D'mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys (CSUR), 2015, 47(3): 1-36.
[4] Snoek C G M, Worrying M, Smeulders A W M. Early versus late fusion in semantic video analysis[C].Proceedings of the 13th annual ACM international conference on Multimedia. Hilton, Singapore, 2005: 399-402.
[5] Ramirez G A, Baltrušaitis T, Morency L P. Modeling latent discriminative dynamic of multidimensional affective signals[C].International Conference on Affective Computing and Intelligent Interaction. Berlin, Heidelberg, 2011: 396-406.
[6] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion[C].Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Berlin, Heidelberg, 2014: 153-162.
[7] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: Metaphor identification with visual features[C].Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, 2016: 160-170.
[8] Hildiegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In The IEEE International Conference on Computer Vision (ICCV), 2011.
[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild arXiv preprint arXiv:1212.0402, 2012.