Video captioning is the task that integrates natural language processing and computer together. For typical approaches, they are based on CNN-RNN, which use the pre-trained Convolutional Neural Network (CNN) to extract image feature and the Recurrent Neural Network (RNN) to generate captions word by word. However, most of the approaches only use the global video feature and lose the spatial and motion information. To address the aforementioned problem, a novel video captioning method based on multi-feature fusion is proposed. This method extracts the spatial features, motion features and video features of each frame, and all the features are fused to generate video captions. The fused features are input into long-term and short-term memory (LSTM) which is used as the natural language generating module. Multiple natural language modules are trained by different feature combinations, and then fused in later stages. First, one model is selected to obtain multiple possible outputs of the current input, and then the probabilities of the current output are calculated by other models. Then the probabilities of these outputs are weighted and the highest probabilities are taken as outputs. In this method, feature fusion methods include pre-fusion and post-fusion. Experiments on the standard test set MSVD
show that the fusion of different types of feature methods can achieve higher evaluation scores; the same type of feature fusion evaluation results will not be higher than a single feature score; the use of features to fine-tune the pre-trained model is not effective. The METEOR score is 0.302, which is 1.34% higher than the current maximum value. It shows that this method can improve the accuracy of automatic video description.

References


**Index Terms**

Computer Science  
Image Processing

**Keywords**

Feature Fusion, Video Captioning, Deep Learning, LSTM