

ABSTRACT

MULTIMODAL HETEROGENEOUS FEATURE FUSION-BASED COMPACT VIDEO EVENT DESCRIPTION METHOD

The present invention relates to the field of computer vision, and in particular, to a video event description method based on multi-modal heterogeneous feature fusion. The method of the present invention includes: clipping videos and extracting dynamic visual features by utilizing an I3D convolutional network, and extracting audio prosody features by utilizing a VGGish model; semantically representing scene object information, generating a scene graph, obtaining an entity code, an attribute code and a relationship code, and performing graph embedding on feature vectors through graph convolution; performing triple multi-modal loop fusion on the three extracted features; performing adaptive multi-modal data balancing so that the dynamic visual feature and the audio prosody feature fit with each other to ensure reasonable event extraction; and describing a reconstruction decoder, detecting a video event by means of a description reconstruction algorithm, and generating a description of a video scene event according to a pre-trained language dictionary. With the method, the problem that meaningless descriptions such as co-occurrence descriptions are frequently generated in a video description method is effectively solved, and the relationship of scene events is mined by effectively utilizing multi-modal information.

<Fig.1>