Claims

1. A video action recognition method, comprising:

obtaining a video, the video comprises multiple consecutive video frames;

inputting the video into a hidden layer of a neural network model for processing, so as to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the hidden layer of the neural network model comprises a plurality of processing units, and;

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequentially identifying spatial features and temporal features of the input video, comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

based on the spatial semantics, performing 1D convolution on the video on time axis to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the concatenating the spatial features and the temporal features of the input video in parallel, comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the

temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

2. The method according to claims 1, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

3. The method according to claim 2, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

4. A video action recognition apparatus, comprising:

obtaining module, obtaining a video, the video comprises multiple consecutive video frames;

processing module, processing the input video based a neural network model to obtain a

recognition object and a motion corresponding to the recognition object of the input video, wherein the processing module comprises a plurality of processing units and;

outputting module, outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, comprising:

sequential processing sub-module, identifying spatial features and temporal features of the input video as input of the at least one processing unit;

parallel processing sub-module, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation; and

outputting sub-module, outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequential processing sub-module comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

performing 1D convolution on the video on time axis based on the spatial semantics to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the parallel processing sub-module comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

5. The apparatus according to claims 4, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

6. The apparatus according to claim 5, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

7. A video action recognition device, comprising:

a memory, configured to store instructions executed by one or more processors of a system, and

a processor, being one of the processors of the system, and configured to execute the instructions to implement the video action recognition method，wherein：

obtaining a video, the video comprises multiple consecutive video frames;

inputting the video into a hidden layer of a neural network model for processing, so as to obtain a recognition object and a motion corresponding to the recognition object of the input video, wherein the hidden layer of the neural network model comprises a plurality of processing units, and;

outputting an action recognition result of the input video based on the recognition object in the video and the motion corresponding to the recognition object;

wherein

in at least one processing unit, sequentially identifying spatial features and temporal features of the input video as input of the at least one processing unit, concatenating the spatial features and the temporal features of the input video in parallel, and performing a pointwise convolution operation, and outputting spatial semantics and temporal semantics of the input video, wherein the spatial semantics are used to represent the recognition object, and the temporal semantics are used to represent the motion corresponding to the recognition object;

wherein the sequentially identifying spatial features and temporal features of the input video, comprises:

performing 2D convolution on the video on spatial axis to obtain the spatial features, to learn the spatial semantics of the input video; and

based on the spatial semantics, performing 1D convolution on the video on time axis to obtain the temporal features, to learn the temporal semantics of the input video;

wherein the concatenating the spatial features and the temporal features of the input video in parallel, comprises:

accumulating channels over the total number of channels of the spatial feature and the channels of the temporal feature, to obtain a concatenated feature map, the spatial feature and the temporal feature are independent of each other; and

calculating all channels of the concatenated feature map to obtain values of the concatenated feature map, so as to integrate the spatial and temporal features.

8. The device according to claims 7, wherein the neural network model comprises an input layer, multiple hidden layers, and an output layer,

the hidden layer further comprises a first convolutional layer, a fully convolutional layer, and a global average pooling layer, and

the fully convolutional layer comprises at least one processing unit,

wherein the consecutive video frames pass through the first convolutional layer that comprises 24 or 32 output channels and on which three-dimensional convolution operation is performed, to obtain 24 or 32 output channels of consecutive feature maps of the consecutive video frames;

the consecutive feature maps of the consecutive video frames are processed by the fully convolutional layer to obtain the recognition object of the input video and the motion corresponding to the recognition object; and

the recognition object of the input video and the motion corresponding to the recognition object are processed by the global average pooling layer and are pooled as the action recognition result of the input video.

9. The device according to claim 8, wherein the neural network model is trained based on something-something v1 and v2 datasets for which pre-training is not needed.

10. A computer-readable storage medium encoded with a computer program, wherein instructions are stored on the computer-readable medium, and when the instructions are executed on a computer, the computer performs the video action recognition method according to any one of claims 1-3.