

Claims

1. An online temporal action detection method, comprising:

obtaining a video online and extracting an initial feature vector including a temporal sequence structure for action detection from the video;

performing candidate action proposal generation processing, constructing a candidate action proposal according to the initial feature vector, and generating a confidence score associated with the candidate action proposal, wherein the confidence score represents a degree of matching between the candidate action proposal and the ground truth action proposal in the video; and

performing action proposal filtering processing, to filter out a redundant action proposal from the candidate action proposal, and output a generated action proposal;

wherein in the performing the generation processing on a candidate action proposal:

performing an online one-dimensional convolution operation for at least once based on the initial feature vector, and outputting a basic feature vector one time point by one time point; and

respectively performing temporal evaluation processing and proposal evaluation processing based on the basic feature vector, wherein

in the temporal evaluation processing, receiving a single basic feature vector one time point by one time point, and performing an operation comprising a recurrent neural network, to generate a starting point probability sequence and an ending point probability sequence of the candidate action proposal one time point by one time point;

in the proposal evaluation processing, cumulatively receiving a predetermined number of basic feature vectors generated one time point by one time point, performing an operation comprising sampling and convolution, and outputting a classification confidence vector and a regression confidence vector; and

calculating the confidence score based on the starting point probability sequence, the ending point probability sequence, the classification confidence vector, and the regression confidence vector; wherein in the online one-dimensional convolution operation, inputting the initial feature vector to a first feature buffer one time point by one time point, wherein the first feature buffer can store k initial feature vectors arranged one time point by one time point, and k is a size of a convolution kernel of the online one-dimensional convolution operation.

2. The method according to claim 1, wherein inputting the initial feature vector to a first feature buffer one time point by one time point further comprises,

inputting the initial feature vector into the first feature buffer in online one-dimensional

convolution one time point by one time point, and deleting the initial feature vector first inputted into the first feature buffer, to ensure that k initial feature vectors inputted most recently one time point by one time point are stored in the first feature buffer, wherein k represents the size of the convolution kernel; and

the online one-dimensional convolution operation further comprises,

performing a convolution operation on the k initial feature vectors in the first feature buffer by using the convolutional kernel, and obtaining the basic feature vector as an output result.

3. The method according to claim 1, wherein the temporal evaluation processing comprises:

receiving a single basic feature vector one time point by one time point, and performing an operation comprising a recurrent neural network, wherein processing on the basic feature vector at a current time point by the recurrent neural network combines a result of processing on the basic feature vector at a previous time point.

4. The method according to claim 3, wherein the recurrent neural network further comprises, the recurrent neural network is a long short-term memory network;

receiving, by the long short-term memory network, a single basic feature vector, an original hidden state h_{t-1} and an original cell state c_{t-1} obtained from a long short-term memory network of a previous time point, and calculating and outputting a current new hidden state h_t based on the basic feature vector, the original hidden state h_{t-1} and the original cell state c_{t-1} ; and

the temporal evaluation processing further comprises,

inputting the new hidden state h_t into a fully connected layer one time point by one time point, and outputting the starting point probability sequence and the ending point probability sequence by using an activation function, wherein the starting point probability sequence consists of a probability $p_{s,t}$ that a time point t is used as a starting point of the candidate action proposal, and the ending point probability sequence consists of a probability $p_{e,t}$ that the time point t is used as an ending point of the candidate action proposal.

5. The method according to claim 4, wherein the method further comprises setting an initial hidden state h_0 and an initial cell state c_0 to be 0.

6. The method according to claim 1, wherein in the proposal evaluation processing, inputting the basic feature vector to a second feature buffer one time point by one time point, wherein the second feature buffer cumulatively receives and stores a predetermined number of basic feature vectors one time point by one time point.

7. The method according to claim 6, wherein inputting the basic feature vector to a second feature buffer one time point by one time point further comprises:

accumulatively receiving, by the second feature buffer, the predetermined number of basic

feature vectors generated one time point by one time point, and storing the basic feature vectors as a basic feature sequence $S_F \in \mathbb{R}^{C \times D}$;

the proposal evaluation processing further comprises,

inputting the basic feature sequence and a sampling weight matrix $W \in \mathbb{R}^{N \times D}$ to a boundary matching layer to perform point multiplication, and obtaining and outputting a boundary matching matrix $M \in \mathbb{R}^{C \times N}$ having a sampling duration N and a time region length from 1 to D , and using the time point t as an ending point; and

inputting the boundary matching matrix into a convolutional neural network including at least one layer to perform convolution operation, and outputting the classification confidence vector $M_{cc} \in \mathbb{R}^D$ and the regression confidence vector $M_{cr} \in \mathbb{R}^D$;

wherein C denotes the number of channels of the basic feature vector, D denotes the length of the longest candidate action proposal in the basic feature sequence, the sampling weight matrix denotes the weight of the basic feature sequence S_F in the sampling duration N , M_{cc} denotes the probability of whether the candidate action proposal having a length from 1 to D and using a current time as an ending point is an action proposal, and M_{cr} denotes the probability that the candidate action proposal having a length from 1 to D and using a current time as an ending point is close to an ground truth action proposal.

8. The method according to claim 7, wherein the at least one layer of convolutional neural network may comprise four layers of convolutional layers, wherein

the first layer is a two-dimensional convolutional layer, and acts on two dimensions of the sampling duration N and the length D of the boundary matching matrix to integrate information of sampling points;

both the second layer and the fourth layer are one-dimensional convolution layers, and act on a dimension of the length D to integrate information of channels; and

the third layer is a one-dimensional convolutional layer, and acts on a dimension of the length D to integrate information of lengths of adjacent proposals.

9. The method according to claim 1, wherein the constructing a candidate action proposal further comprises:

determining that the time point t is a candidate action starting point, and adding the time point t of the candidate action starting point to a list of candidate action starting points;

determining that the time point t is a candidate action ending point, selecting, from the list of candidate action starting points, the candidate starting point that can form the candidate action proposal having a length smaller than a maximum length D with the time point t , and outputting a

set of candidate action proposals $\Phi_c = \{\phi_i = [t_{s,i}, t]\}_{i=1}^{N_c}$;

wherein ϕ_i denotes an i^{th} candidate action proposal, N_c denotes the number of candidate action proposals, and $t_{s,i}$ denotes a start time of the i^{th} candidate action proposal.

10. The method according to claim 1, wherein the confidence score is calculated by the following formula:

$$\text{score}_i = p_{s,t_{s,i}} \times p_{e,t} \times \sqrt{M_{cc,t}[t - t_{s,i} + 1] \times M_{cr,t}[t - t_{s,i} + 1]},$$

wherein $t - t_{s,i} + 1$ denotes a length of the candidate action proposal ϕ_i , $p_{s,t}$ and $p_{e,t}$ respectively denote probabilities that the time point t is used as a starting point and an ending point of the candidate action proposal, $M_{cc,t}$ denotes the probability of whether the candidate action proposal having a length from 1 to D and using a time point t as an ending point is an action proposal, and $M_{cr,t}$ denotes the probability that the candidate action proposal having a length from 1 to D and using a time point t as an ending point is close to the ground truth action proposal.

11. An online temporal action detection apparatus, comprising:

a video obtaining unit, configured to obtain a video online and extracting an initial feature vector including a temporal sequence structure for action detection from the video;

a candidate action proposal generation unit, configured to construct a candidate action proposal based on the initial feature vector, and generate a confidence score associated with the candidate action proposal, wherein the confidence score represents a degree of matching between the candidate action proposal and the ground truth action proposal in the video; and

an action proposal filtering unit, configured to filter out a redundant action proposal from the candidate action proposal and output a generated action proposal;

wherein in the candidate action proposal generating unit:

an online one-dimensional convolution operation subunit is configured to perform an online one-dimensional convolution operation for at least once based on the initial feature vector, and output a basic feature vector one time point by one time point; and

respectively input the basic feature vector into a temporal evaluation subunit and a proposal evaluation subunit to perform an operation, wherein

the temporal evaluation subunit is configured to receive a single basic feature vector one

time point by one time point, and perform an operation comprising a recurrent neural network, to generate a starting point probability sequence and an ending point probability sequence of the candidate action proposal one time point by one time point;

the proposal evaluation subunit is configured to cumulatively receive a predetermined number of basic feature vectors generated one time point by one time point, perform an operation comprising sampling and convolution, and output a classification confidence vector and a regression confidence vector; and

a confidence score calculation subunit is configured to calculate the confidence score based on the starting point probability sequence, the ending point probability sequence, the classification confidence vector, and the regression confidence vector; wherein in the online one-dimensional convolution operation subunit, inputting the initial feature vector to a first feature buffer one time point by one time point, wherein the first feature buffer can store k initial feature vectors arranged one time point by one time point, and k is a size of a convolution kernel of the online one-dimensional convolution operation.

12. The apparatus according to claim 11, wherein the online one-dimensional convolution operation subunit further comprises:

the first feature buffer, used to receive the initial feature vector one time point by one time point and delete the initial feature vector first inputted into the first feature buffer, to ensure that k initial feature vectors most recently inputted one time point by one time point are stored in the first feature buffer, wherein k represents the size of the convolution kernel; and

a one-dimensional convolution layer, used to perform a convolution operation on the k initial feature vectors in the first feature buffer, and obtain the basic feature vector as an output result.

13. The apparatus according to claim 11, wherein the temporal evaluation subunit comprises: receiving a single basic feature vector one time point by one time point and performing an operation comprising a recurrent neural network, wherein processing on the basic feature vector at a current time point by the recurrent neural network combines a result of processing on the basic feature vector at a previous time point.

14. The method according to claim 13, wherein the temporal evaluation subunit further comprises:

a long short-term memory network, the recurrent neural network being a long short-term memory network, used to receive a single basic feature vector, an original hidden state h_{t-1} and an

original cell state c_{t-1} obtained from a long short-term memory network of a previous time point, and calculating and outputting a current new hidden state h_t based on the basic feature vector, the original hidden state h_{t-1} and the original cell state c_{t-1} ; and

a fully connected layer, used to receive the new hidden state h_t one time point by one time point, and output the starting point probability sequence and the ending point probability sequence by using an activation function, wherein the starting point probability sequence consists of a probability $p_{s,t}$ that a time point t is used as a starting point of the candidate action proposal, and the ending point probability sequence consists of a probability $p_{e,t}$ that the time point t is used as an ending point of the candidate action proposal.

15. The apparatus according to claim 11, wherein in the proposal evaluation subunit, inputting the basic feature vector into a second feature buffer one time point by one time point, wherein the second feature buffer cumulatively receives and stores a predetermined number of basic feature vectors one time point by one time point.

16. The apparatus according to claim 15, wherein the proposal evaluation subunit further comprises:

the second feature buffer, used to accumulatively receive a predetermined number of basic feature vectors generated one time point by one time point and store the basic feature vectors as a basic feature sequence $S_F \in \mathbb{R}^{C \times D}$;

a boundary matching layer, used to perform point multiplication on the basic feature sequence and a sampling weight matrix $W \in \mathbb{R}^{N \times D}$, and obtain and output a boundary matching matrix $M \in \mathbb{R}^{C \times N}$ having a sampling duration N and a time region length from 1 to D , and using the time point t as an ending point; and

a convolutional neural network, the convolutional neural network comprising at least one layer, used to input the boundary matching matrix to perform convolution operation, and output the classification confidence vector $M_{cc} \in \mathbb{R}^D$ and the regression confidence vector $M_{cr} \in \mathbb{R}^D$;

wherein C denotes the number of channels of the basic feature vector, D denotes the length of the longest candidate action proposal in the basic feature sequence, the sampling weight matrix denotes the weight of the basic feature sequence S_F in the sampling duration N , M_{cc} denotes the probability of whether the candidate action proposal having a length from 1 to D and using a current time as an ending point is an action proposal, and M_{cr} denotes the probability that the candidate action proposal having a length from 1 to D and using a current time as an ending point is close to an ground truth action proposal.

17. An online temporal action detection device, comprising:

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

a processor, being one of the processors of the system, configured to execute the instruction to implement the temporal action detection method according to any one of claims 1 to 10.

18. A computer readable storage medium encoded with a computer program, wherein an instruction is stored on the computer readable medium, and when being executed on a computer, the instruction enables the computer to perform the temporal action detection method according to any one of claims 1 to 10.