

CLAIMS

1. A time domain action detecting method, comprising:

obtaining a time domain interval in a video with an action instance and at least one adjacent segment in the time domain interval;

separately extracting action features of at least two video segments in candidate segments, wherein the candidate segments comprise a video segment corresponding to the time domain interval and the adjacent segments thereof;

pooling the action features of the at least two video segments in the candidate segments, to obtain a global feature of the video segment corresponding to the time domain interval; and

determining, based on the global feature, an action integrity score of the video segment corresponding to the time domain interval.

2. The method according to claim 1, wherein the at least one adjacent segment comprises: a first adjacent segment in the video with a time sequence located in front of the time domain interval, and/or a second adjacent segment in the video with a time sequence located behind the time domain interval; and the first adjacent segment and the second segment respectively comprise at least one video segment.

3. The method according to claim 1 or 2, wherein the obtaining a time domain interval in a video with an action instance and at least one adjacent segment in the time domain interval comprises:

performing actionness estimation separately on at least one video segment in the video, to obtain a time sequence actionness sequence;

performing action position prediction based on the time sequence actionness sequence, to obtain the time domain interval in the video with an action instance, the time domain interval comprising a start time and an end time; and

extracting, from the video, the first adjacent segment before the time domain interval and/or the second adjacent segment after the time domain interval.

4. The method according to claim 3, wherein the performing actionness estimation separately

on at least one video segment in the video, to obtain a time sequence actionness sequence comprises:

for any video segment in the video separately: extracting an image frame as an original image, and performing actionness estimation on the original image, to obtain a first actionness value; extracting a light stream of the any video segment, merging obtained light stream field pictures, to obtain a spliced light stream field image, and performing actionness estimation on the spliced light stream field image, to obtain a second actionness value;

obtaining an actionness value of the any video segment from the first actionness value and the second actionness value; and

forming the time sequence actionness sequence by the actionness values of all video segments based on a time sequence relation.

5. The method according to claim 4, wherein after the obtaining the actionness value of any video segment, the method further comprises:

normalizing the actionness value of the any video segment, to obtain a normalized actionness value; and

the time sequence actionness sequence comprising: a time sequence actionness sequence formed by the normalized actionness value.

6. The method according to any one of claims 1 to 5, further comprising:

obtaining, based on the action feature of the video segment corresponding to the time domain interval, a category score of at least one action category of the video segment corresponding to the time domain interval; and

determining, according to the category score of the at least one action category of the video segment corresponding to the time domain interval, a detected action category of the video segment corresponding to the time domain interval.

7. The method according to claim 6, further comprising:

outputting the time domain interval and the detected action category of the video segment corresponding to the time domain interval.

8. The method according to claim 6 or 7, wherein the obtaining, based on an action feature of the video segment corresponding to the time domain interval, a category score of at least one action category of the video segment corresponding to the time domain interval comprises:

separately obtaining, based on the action feature of the at least one video segment corresponding to the time domain interval, a score of the at least one video segment corresponding to the time domain interval separately belonging to the at least action category; and

summing scores of the at least one video segment corresponding to the time domain interval separately belonging to the same action category, to obtain the category score of the at least one action category of the video segment corresponding to the time domain interval.

9. The method according to any one of claims 1 to 8, wherein the pooling the action features of the at least two video segments in the candidate segments comprises:

performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments.

10. The method according to claim 9, wherein after the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments, the method further comprises:

merging pooling features obtained after the time domain pyramid-typed pooling.

11. The method according to claim 10, wherein before the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments, the method further comprises:

presetting a value of the number K of pooling layers to be 1;

the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments comprising:

for any first to-be-pooled segment with a value of a preset partition part number B_K to be 1, obtaining the pooling feature of the any first to-be-pooled segment from the action feature of the at least one video segment in the any first to-be-pooled segment;

for any second to-be-pooled segment with the value of the preset partition part number B_K to

be greater than 1, segmenting all video segments in the any second to-be-pooled segment into B_K parts, obtaining the pooling feature of a corresponding part separately from the action features of each part of the video segments in the B_K parts, and merging the pooling features of the B_K parts, to obtain the pooling feature of the any second to-be-pooled segment; and

the first to-be-pooled segment comprising the video segment corresponding to the time domain interval, any one or more of the first adjacent segment and the second adjacent segment; the second to-be-pooled segment comprising other to-be-pooled segments in the candidate segments except the first to-be-pooled segment.

12. The method according to claim 10, wherein before the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments, the method further comprises:

presetting a value of the number K of pooling layers to be greater than 1;

the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments comprising:

separately for the k^{th} pooling layer: for any first to-be-pooled segment with a value of a preset partition part number B_K to be 1, obtaining the pooling feature of the any first to-be-pooled segment at the k^{th} layer from the action feature of the at least one video segment in the any first to-be-pooled segment; for any second to-be-pooled segment with the value of the preset partition part number B_K to be greater than 1, segmenting all video segments in the any second to-be-pooled segment into B_K parts, obtaining the pooling feature of a corresponding part separately from the action features of each part of the video segments in the B_K parts, and merging the pooling features of the B_K parts, to obtain the pooling feature of the any second to-be-pooled segment at the k^{th} layer;

separately for at least one to-be-pooled segment: merging the pooling features of K pooling layers at a preset manner, to obtain the pooling feature of the at least one to-be-pooled segment; and

the first to-be-pooled segment comprising the video segment corresponding to the time domain interval, any one or more of the first adjacent segment and the second adjacent

segment; the second to-be-pooled segment comprising other to-be-pooled segments in the candidate segments except the first to-be-pooled segment.

13. The method according to any one of claims 10 to 12, wherein the merging pooling features obtained after the time domain pyramid-typed pooling, to obtain a global feature of the video segment corresponding to the time domain interval comprises:

separately merging the pooling feature of the first adjacent segment, the pooling feature of the video segment corresponding to the time domain interval, and the pooling feature of the second adjacent segment, to obtain the global feature of the video segment corresponding to the time domain interval upon time sequence structuration.

14. The method according to claim 10, wherein before the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments, the method further comprises:

presetting a value of the number K of pooling layers to be greater than 1;

the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments comprising:

separately for the k^{th} pooling layer: identifying whether the value of the preset partition part number B_K is 1;

if the value of the preset partition part number B_K is 1, obtaining the pooling features of the candidate segments at the k^{th} layer from the action feature of the at least one video segment in the candidate segments; and

if the value of the preset partition part number B_K is greater than 1, segmenting all video segments in the candidate segments into B_K parts, obtaining the pooling feature of a corresponding part separately from the action features of each part of the video segments in the B_K parts, and merging the pooling features of the B_K parts, to obtain the pooling features of the candidate segments at the k^{th} layer.

15. The method according claim 14, wherein the merging pooling features obtained after the time domain pyramid-typed pooling, to obtain a global feature of the video segment corresponding

to the time domain interval comprises:

merging the pooling features of K pooling layers at a preset manner, to obtain the global feature of the video segment corresponding to the time domain interval upon time sequence structuration.

16. The method according to claim 10, wherein before the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments, the method further comprises:

presetting a value of the number K of pooling layers to be 1;

the performing time domain pyramid-typed pooling processing on the action features of the at least two video segments in the candidate segments comprising:

identifying whether the value of the preset partition part number B_K is 1;

if the value of the preset partition part number B_K is 1, separately for any to-be-pooled segment in at least two to-be-pooled segments, obtaining the pooling feature of the any to-be-pooled segment from the action features of the at least two video segments in the any to-be-pooled segment; the to-be-pooled segment comprising the video segments in the time domain interval, the first adjacent segment, and the second adjacent segment; merging the pooling features of all to-be-pooled segments, to obtain the global feature of the video segment corresponding to the time domain interval upon time sequence structuration.

if the value of the preset partition part number B_K is greater than 1, segmenting all video segments in the candidate segments into B_K parts, obtaining the pooling feature of a corresponding part separately from the action features of each part of the video segments in the B_K parts, and merging the pooling features of the B_K parts, to obtain the global feature of the video segment corresponding to the time domain interval upon time sequence structuration.

17. The method according to any one of claims 1 to 16, wherein the determining, based on the global feature, an action integrity score of the video segment corresponding to the time domain interval comprises:

classifying the global feature, to obtain at least one action category corresponding to the

candidate segments; and

separately identifying, based on the global feature, the integrity degree of an action instance in a current action category comprised in the video segment corresponding to the time domain interval by means of a classifier related to at least one action category corresponding to the candidate segments, to obtain the action integrity score of the video segment corresponding to the time domain interval separately corresponding to the at least one action category.

18. The method according to any one of claims 1 to 17, further comprising:

determining, based on the category score and the action integrity score of the video segment corresponding to the time domain interval, whether the video segment corresponding to the time domain interval comprises an integral action instance corresponding to the at least one action category.

19. The method according to claim 18, wherein the determining, based on the category score and the action integrity score of the video segment corresponding to the time domain interval, whether the video segment corresponding to the time domain interval comprises an integral action instance corresponding to the at least one action category comprises:

separately for any action category corresponding to the video segment corresponding to the time domain interval, calculating a product or a sum of the category score and the action integrity score of the video segment corresponding to the time domain interval corresponding to the any action category, to obtain a confidence value of whether the video segment corresponding to the time domain interval comprises the integral action instance in the any action category; and

determining, based on whether the confidence value is greater than a preset threshold, whether the video segment corresponding to the time domain interval comprises the integral action instance in the any action category.

20. The method according to claim 19, further comprising:

if the number of the action categories corresponding to the integral action instance comprised in the video segment corresponding to the time domain interval is greater than 1, selecting one action category corresponding to a maximum confidence value as a detected action category of the video segment corresponding to the time domain interval.

21. The method according to claim 18, wherein the determining, based on the category score and the action integrity score of the video segment corresponding to the time domain interval, whether the video segment corresponding to the time domain interval comprises an integral action instance corresponding to the at least one action category comprises:

identifying whether a preset action category with a category score greater than a preset first numeral value exists in the category scores of at least one action category of the video segment corresponding to the time domain interval; and identifying whether a preset action category with an action integrity score greater than a preset second numeral value exists in the action integrity scores;

in response to existence of the preset action category with the category score greater than the preset first numeral value and/or the preset action category with the action integrity score greater than the preset second numeral value, using the preset action category with the category score greater than the preset first numeral value and/or the preset action category with the action integrity score greater than the preset second numeral value as a target action category, and determining, based on the category score and the action integrity score of the target action category corresponding to the video segment corresponding to the time domain interval, whether the video segment corresponding to the time domain interval comprises an integral action instance in the target action category; and

when outputting the time domain interval and the detected action category of the video segment corresponding to the time domain interval, outputting the target action category and the time domain interval thereof.

22. The method according to any one of claims 18 to 21, further comprising:

if the video segment corresponding to the time domain interval comprises the integral action instance of at least one action category, performing position regression on the time domain interval based on the global feature, to obtain a start-end position offset amount, the start-end position offset amount comprising a position offset amount between a start time and an end time in the time domain interval;

adjusting the time domain interval by means of the start-end position offset amount; and

when outputting the time domain interval and the detected action category of the video segment corresponding to the time domain interval, outputting the adjusted time domain interval and the detected action category of the video segment corresponding to the time domain interval.

23. An electronic device, comprising a processor and a memory;

the memory, configured to store at least one executable instruction, and the executable instruction enabling the processor to execute corresponding operations of the time domain action detecting method according to any one of claims 1 to 22.

24. A computer storage medium, configured to store computer-readable instructions, wherein when the instructions are executed, operations in steps of the time domain action detecting method according to any one of claims 1 to 22 are implemented.