

CLAIMS

What is claimed is:

1. A video processing method, comprising:

5 performing a feature extraction on a plurality of target video frames of a video to be processed through a feature extraction network, to obtain feature maps of the plurality of target video frames;

performing an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network, to obtain action recognition features of the plurality of target video frames; wherein M is an integer greater than or equal to 1, the action recognition process comprises a spatiotemporal feature extraction process based on the feature maps of the plurality of target video frames, and a motion feature extraction process based on motion difference information between the feature maps of the plurality of target video frames, the action recognition feature comprises spatiotemporal feature information and motion feature information; and

15 determining a classification result of the video to be processed according to the action recognition features of the plurality of target video frames.

2. The method according to claim 1, wherein performing an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network to obtain action recognition features of the plurality of target video frames comprises: processing the feature maps of the plurality of target video frames through a first level action recognition network, to obtain first level action recognition features;

20 processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features, wherein i is an integer and $1 < i < M$; action recognition features of respective levels correspond to the feature maps of the plurality of target video frames; and

processing $(M-1)^{\text{th}}$ level action recognition features through an M^{th} level action recognition network, to obtain action recognition features of the plurality of target video frames.

30

3. The method according to claim 2, wherein processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features comprises:

performing a first convolution process on the $(i-1)^{\text{th}}$ level action recognition features to obtain first feature information, wherein the first feature information corresponds to the feature maps

- of the plurality of target video frames respectively;
performing a spatiotemporal feature extraction process on the first feature information, to
obtain spatiotemporal feature information;
performing a motion feature extraction process on the first feature information, to obtain
5 motion feature information; and
obtaining the i^{th} level action recognition features at least based on the spatiotemporal feature
information and the motion feature information.
4. The method according to claim 3, wherein obtaining the i^{th} level action recognition features
10 at least based on the spatiotemporal feature information and the motion feature information
comprises:
obtaining the i^{th} level action recognition features based on the spatiotemporal feature
information, the motion feature information, and the $(i-1)^{\text{th}}$ level action recognition features.
- 15 5. The method according to claim 3, wherein performing a spatiotemporal feature extraction
process on the first feature information to obtain spatiotemporal feature information comprises:
performing dimensional reconstruction processes respectively on the first feature information
corresponding to the feature maps of the plurality of target video frames, to obtain second
feature information, wherein the second feature information has a different dimension from the
20 first feature information;
performing second convolution processes respectively on channels of the second feature
information to obtain third feature information, wherein the third feature information represents
time features of the feature maps of the plurality of target video frames;
performing a dimensional reconstruction process on the third feature information to obtain
25 fourth feature information, wherein the fourth feature information has a same dimension as the
first feature information; and
performing a spatial feature extraction process on the fourth feature information, to obtain the
spatiotemporal feature information.
- 30 6. The method according to claim 5, wherein, the first feature information comprises multiple
row vectors or column vectors, and
performing dimensional reconstruction processes respectively on the first feature information
corresponding to the feature maps of the plurality of target video frames comprises:
performing splicing processes on the multiple row vectors or column vectors of the first feature
35 information to obtain the second feature information, wherein the second feature information

comprises one row vector or column vector.

7. The method according to any one of claims 3 to 6, wherein performing a motion feature extraction process on the first feature information to obtain motion feature information comprises:

performing dimensional reduction processes on channels of the first feature information to obtain fifth feature information, wherein the fifth feature information corresponds to respective target video frames of the video to be processed;

performing a third convolution process on the fifth feature information corresponding to a $(k+1)^{\text{th}}$ target video frame, and subtracting it by the fifth feature information corresponding to a k^{th} target video frame, to obtain sixth feature information corresponding to the k^{th} target video frame, where k is an integer and $1 \leq k < T$, T is the number of the target video frames, and T is an integer greater than 1, the sixth feature information represents motion difference information between the fifth feature information corresponding to the $(k+1)^{\text{th}}$ target video frame and the fifth feature information corresponding to the k^{th} target video frame; and performing a feature extraction process on the sixth feature information corresponding to the respective target video frames, to obtain the motion feature information.

8. The method according to any one of claims 4 to 7, wherein obtaining the i^{th} level action recognition features based on the spatiotemporal feature information, the motion feature information, and the $(i-1)^{\text{th}}$ level action recognition features comprises:

performing a summation process on the spatiotemporal feature information and the motion feature information, to obtain seventh feature information; and

performing on the seventh feature information a fourth convolution process, and a summation process with the $(i-1)^{\text{th}}$ level action recognition features, to obtain the i^{th} level action recognition features.

9. The method according to any one of claims 1 to 8, wherein determining a classification result of the video to be processed according to the action recognition features of the plurality of target video frames comprises:

performing a full connection process on the action recognition features of the target video frames respectively, to obtain classification information of the respective target video frames; and

performing an averaging process on the classification information of the respective target video frames, to obtain the classification result of the video to be processed.

10. The method according to any one of claims 1 to 9, further comprising:
determining a plurality of target video frames from the video to be processed.
- 5 11. The method according to claim 10, wherein determining a plurality of target video frames from the video to be processed comprises:
dividing the video to be processed into a plurality of video clips; and
determining randomly at least one target video frame from each video clip to obtain the plurality of target video frames.
- 10 12. The method according to any one of claims 1 to 11, wherein the video processing method is implemented through a neural network, and the neural network at least comprises the feature extraction network and the M-level action recognition network,
the method further comprises:
15 training the neural network by a sample video and a category label of the sample video.
13. The method according to claim 12, wherein training the neural network by a sample video and a category label of the sample video comprises:
determining a plurality of sample video frames from the sample video,
20 processing the sample video frames through the neural network, to determine a classification result of the sample video;
determining a network loss of the neural network according to the classification result and a category label of the sample video; and
adjusting network parameters of the neural network according to the network loss.
- 25 14. A video processing apparatus, comprising:
a feature extraction module configured to perform a feature extraction on a plurality of target video frames of a video to be processed through a feature extraction network, to obtain feature maps of the plurality of target video frames;
30 an action recognition module configured to perform an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network, to obtain action recognition features of the plurality of target video frames; wherein M is an integer greater than or equal to 1, the action recognition process comprises a spatiotemporal feature extraction process based on the feature maps of the plurality of target
35 video frames, and a motion feature extraction process based on motion difference information

between the feature maps of the plurality of target video frames, the action recognition feature comprises spatiotemporal feature information and motion feature information; and
a classification module configured to determine a classification result of the video to be processed according to the action recognition features of the plurality of target video frames.

5

15. The apparatus according to claim 14, wherein the action recognition module is further configured to:

process the feature maps of the plurality of target video frames through a first level action recognition network, to obtain first level action recognition features;

10 process $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features, wherein i is an integer and $1 < i < M$, wherein action recognition features of respective levels correspond to the feature maps of the plurality of target video frames; and

15 process $(M-1)^{\text{th}}$ level action recognition features through an M^{th} level action recognition network, to obtain action recognition features of the plurality of target video frames.

16. The apparatus according to claim 15, wherein the action recognition module is further configured to:

20 perform a first convolution process on the $(i-1)^{\text{th}}$ level action recognition features to obtain first feature information, wherein the first feature information corresponds to the feature maps of the plurality of target video frames respectively;

perform a spatiotemporal feature extraction process on the first feature information, to obtain the spatiotemporal feature information;

25 perform a motion feature extraction process on the first feature information, to obtain the motion feature information; and

obtain the i^{th} level action recognition features based on the spatiotemporal feature information and the motion feature information.

30 17. The apparatus according to claim 16, wherein the action recognition module is further configured to:

obtain the i^{th} level action recognition features based on the spatiotemporal feature information, the motion feature information, and the $(i-1)^{\text{th}}$ level action recognition features.

35 18. The apparatus according to claim 16, wherein the action recognition module is further configured to:

perform dimensional reconstruction processes respectively on the first feature information corresponding to the feature maps of the plurality of target video frames, to obtain second feature information, wherein the second feature information has a different dimension from the first feature information;

5 perform second convolution processes respectively on channels of the second feature information to obtain third feature information, wherein the third feature information represents time features of the feature maps of the plurality of target video frames;

perform a dimensional reconstruction process on the third feature information to obtain fourth feature information, wherein the fourth feature information has a same dimension as the first

10 feature information; and

perform a spatial feature extraction process on the fourth feature information, to obtain the spatiotemporal feature information.

19. The apparatus according to claim 18, wherein the first feature information comprises

15 multiple row vectors or column vectors, and

the action recognition module is further configured to:

perform splicing processes on the multiple row vectors or column vectors of the first feature information to obtain the second feature information, wherein the second feature information comprises one row vector or column vector.

20

20. The apparatus according to any one of claims 16 to 19, wherein the action recognition module is further configured to:

perform dimensional reduction processes on channels of the first feature information to obtain fifth feature information, wherein the fifth feature information corresponds to respective target

25 video frames of the video to be processed;

perform a third convolution process on the fifth feature information corresponding to a $(k+1)^{\text{th}}$ target video frame, and subtracting it by the fifth feature information corresponding to a k^{th} target video frame, to obtain sixth feature information corresponding to the k^{th} target video frame, where k is an integer and $1 \leq k < T$, T is the number of target video frames, and T is an

30 integer greater than 1, the sixth feature information represents motion difference information between the fifth feature information corresponding to the $(k+1)^{\text{th}}$ target video frame and the fifth feature information corresponding to the k^{th} target video frame; and

perform a feature extraction process on the sixth feature information corresponding to the respective target video frames, to obtain the motion feature information.

35

21. The apparatus according to any one of claims 17 to 20, wherein the action recognition module is further configured to:
perform a summation process on the spatiotemporal feature information and the motion feature information, to obtain seventh feature information;
5 perform on the seventh feature information a fourth convolution process, and a summation process with the $(i-1)^{\text{th}}$ level action recognition features, to obtain the i^{th} level action recognition features.
22. The apparatus according to any one of claims 14 to 21, wherein the classification module is
10 further configured to:
perform a full connection process on the action recognition features of the target video frames respectively, to obtain classification information of the respective target video frames; and
perform an averaging process on the classification information of the respective target video frames, to obtain the classification result of the video to be processed.
- 15 23. The apparatus according to any one of claims 14 to 22, further comprising:
a determination module configured to determine a plurality of target video frames from the video to be processed.
- 20 24. The apparatus according to claim 23, wherein the determination module is further configured to:
divide the video to be processed into a plurality of video clips; and
determine randomly at least one target video frame from each video clip to obtain the plurality of target video frames.
- 25 25. The apparatus according to any one of claims 14 to 24, wherein the video processing apparatus is implemented through a neural network, and the neural network at least comprises the feature extraction network and the M-level action recognition network,
the apparatus further comprises:
30 a training module configured to train the neural network by a sample video and a category label of the sample video.
26. The apparatus according to claim 25, wherein the training module is further configured to:
determine a plurality of sample video frames from the sample video;
35 process the sample video frames through the neural network, to determine a classification

result of the sample video;

determine a network loss of the neural network according to the classification result and a category label of the sample video; and

adjust network parameters of the neural network according to the network loss.

5

27. An electronic device, comprising:

a processor; and

a memory for storing processor executable instructions;

wherein the processor is configured to invoke instructions stored in the memory to execute the

10 method according to any one of claims 1 to 13.

28. A computer-readable storage medium having computer program instructions stored thereon, wherein the computer program instructions, when executed by a processor, implement the method according to any one of claims 1 to 13.

15

29. A computer program comprising computer-readable codes, wherein a processor in an electronic device is configured to execute the method according to any one of claims 1 to 13 when the computer-readable codes is running on the electronic device.

CLAIMS

1. A video processing method, comprising:

performing a feature extraction on a plurality of target video frames of a video to be processed through a feature extraction network, to obtain feature maps of the plurality of target video frames;

performing an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network, to obtain action recognition features of the plurality of target video frames; wherein M is an integer greater than or equal to 1, the action recognition process comprises a spatiotemporal feature extraction process based on the feature maps of the plurality of target video frames, and a motion feature extraction process based on motion difference information between the feature maps of the plurality of target video frames, the action recognition feature comprises spatiotemporal feature information and motion feature information; and

determining a classification result of the video to be processed according to the action recognition features of the plurality of target video frames,

wherein performing an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network to obtain action recognition features of the plurality of target video frames comprises:

processing the feature maps of the plurality of target video frames through a first level action recognition network, to obtain first level action recognition features;

processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features, wherein i is an integer and $1 < i < M$; action recognition features of respective levels correspond to the feature maps of the plurality of target video frames; and

processing $(M-1)^{\text{th}}$ level action recognition features through an M^{th} level action recognition network, to obtain action recognition features of the plurality of target video frames, and

wherein processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features comprises:

performing a first convolution process on the $(i-1)^{\text{th}}$ level action recognition features to obtain first feature information, wherein the first feature information corresponds to the feature maps of the plurality of target video frames respectively;

performing a spatiotemporal feature extraction process on the first feature information, to obtain spatiotemporal feature information;

performing a motion feature extraction process on the first feature information, to obtain motion feature information; and

obtaining the i^{th} level action recognition features at least based on the spatiotemporal feature information and the motion feature information.

2. The method according to claim 1, wherein obtaining the i^{th} level action recognition features at least based on the spatiotemporal feature information and the motion feature information comprises:

obtaining the i^{th} level action recognition features based on the spatiotemporal feature information, the motion feature information, and the $(i-1)^{\text{th}}$ level action recognition features.

3. The method according to claim 1, wherein performing a spatiotemporal feature extraction process on the first feature information to obtain spatiotemporal feature information comprises:

performing dimensional reconstruction processes respectively on the first feature information corresponding to the feature maps of the plurality of target video frames, to obtain second feature information, wherein the second feature information has a different dimension from the first feature information;

performing second convolution processes respectively on channels of the second feature information to obtain third feature information, wherein the third feature information represents time features of the feature maps of the plurality of target video frames;

performing a dimensional reconstruction process on the third feature information to obtain fourth feature information, wherein the fourth feature information has a same dimension as the first feature information; and

performing a spatial feature extraction process on the fourth feature information, to obtain the spatiotemporal feature information.

4. The method according to claim 3, wherein, the first feature information comprises multiple row vectors or column vectors, and

performing dimensional reconstruction processes respectively on the first feature

information corresponding to the feature maps of the plurality of target video frames comprises:

performing splicing processes on the multiple row vectors or column vectors of the first feature information to obtain the second feature information, wherein the second feature information comprises one row vector or column vector.

5. The method according to any one of claims 1 to 4, wherein performing a motion feature extraction process on the first feature information to obtain motion feature information comprises:

performing dimensional reduction processes on channels of the first feature information to obtain fifth feature information, wherein the fifth feature information corresponds to respective target video frames of the video to be processed;

performing a third convolution process on the fifth feature information corresponding to a $(k+1)^{\text{th}}$ target video frame, and subtracting it by the fifth feature information corresponding to a k^{th} target video frame, to obtain sixth feature information corresponding to the k^{th} target video frame, where k is an integer and $1 \leq k < T$, T is the number of the target video frames, and T is an integer greater than 1, the sixth feature information represents motion difference information between the fifth feature information corresponding to the $(k+1)^{\text{th}}$ target video frame and the fifth feature information corresponding to the k^{th} target video frame; and

performing a feature extraction process on the sixth feature information corresponding to the respective target video frames, to obtain the motion feature information.

6. The method according to any one of claims 2 to 5, wherein obtaining the i^{th} level action recognition features based on the spatiotemporal feature information, the motion feature information, and the $(i-1)^{\text{th}}$ level action recognition features comprises:

performing a summation process on the spatiotemporal feature information and the motion feature information, to obtain seventh feature information; and

performing on the seventh feature information a fourth convolution process, and a summation process with the $(i-1)^{\text{th}}$ level action recognition features, to obtain the i^{th} level action recognition features.

7. The method according to any one of claims 1 to 6, wherein determining a

classification result of the video to be processed according to the action recognition features of the plurality of target video frames comprises:

performing a full connection process on the action recognition features of the target video frames respectively, to obtain classification information of the respective target video frames; and

performing an averaging process on the classification information of the respective target video frames, to obtain the classification result of the video to be processed.

8. The method according to any one of claims 1 to 7, further comprising:
determining a plurality of target video frames from the video to be processed.

9. The method according to claim 8, wherein determining a plurality of target video frames from the video to be processed comprises:

dividing the video to be processed into a plurality of video clips; and

determining randomly at least one target video frame from each video clip to obtain the plurality of target video frames.

10. The method according to any one of claims 1 to 9, wherein the video processing method is implemented through a neural network, and the neural network at least comprises the feature extraction network and the M-level action recognition network,

the method further comprises:

training the neural network by a sample video and a category label of the sample video.

11. The method according to claim 10, wherein training the neural network by a sample video and a category label of the sample video comprises:

determining a plurality of sample video frames from the sample video,

processing the sample video frames through the neural network, to determine a classification result of the sample video;

determining a network loss of the neural network according to the classification result and a category label of the sample video; and

adjusting network parameters of the neural network according to the network

loss.

12. A video processing apparatus, comprising:

a feature extraction module configured to perform a feature extraction on a plurality of target video frames of a video to be processed through a feature extraction network, to obtain feature maps of the plurality of target video frames;

an action recognition module configured to perform an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network, to obtain action recognition features of the plurality of target video frames; wherein M is an integer greater than or equal to 1, the action recognition process comprises a spatiotemporal feature extraction process based on the feature maps of the plurality of target video frames, and a motion feature extraction process based on motion difference information between the feature maps of the plurality of target video frames, the action recognition feature comprises spatiotemporal feature information and motion feature information; and

a classification module configured to determine a classification result of the video to be processed according to the action recognition features of the plurality of target video frames,

wherein performing an action recognition process on the feature maps of the plurality of target video frames through an M-level action recognition network to obtain action recognition features of the plurality of target video frames comprises:

processing the feature maps of the plurality of target video frames through a first level action recognition network, to obtain first level action recognition features;

processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features, wherein i is an integer and $1 < i < M$, wherein action recognition features of respective levels correspond to the feature maps of the plurality of target video frames; and

processing $(M-1)^{\text{th}}$ level action recognition features through an M^{th} level action recognition network, to obtain action recognition features of the plurality of target video frames, and

wherein processing $(i-1)^{\text{th}}$ level action recognition features through an i^{th} level action recognition network to obtain i^{th} level action recognition features comprises:

performing a first convolution process on the $(i-1)^{\text{th}}$ level action recognition features to obtain first feature information, wherein the first feature information

corresponds to the feature maps of the plurality of target video frames respectively;
performing a spatiotemporal feature extraction process on the first feature information, to obtain the spatiotemporal feature information;
performing a motion feature extraction process on the first feature information, to obtain the motion feature information; and
obtaining the i^{th} level action recognition features based on the spatiotemporal feature information and the motion feature information.

13. An electronic device, comprising:
a processor; and
a memory for storing processor executable instructions;
wherein the processor is configured to invoke instructions stored in the memory to execute the method according to any one of claims 1 to 11.

14. A computer-readable storage medium having computer program instructions stored thereon, wherein the computer program instructions, when executed by a processor, implement the method according to any one of claims 1 to 11.

15. A computer program comprising computer-readable codes, wherein a processor in an electronic device is configured to execute the method according to any one of claims 1 to 11 when the computer-readable codes is running on the electronic device.