CLAIMS

1. A method for recognizing a video, comprising:

extracting features of a first key frame in the video;

performing fusion on the features of the first key frame and fusion features of a second key frame in the video to obtain fusion features of the first key frame, wherein a detection sequence of the second key frame in the video precedes that of the first key frame; and

performing detection on the first key frame according to the fusion features of the first key frame to obtain an object detection result of the first key frame.

2. The method according to claim 1, wherein the second key frame comprises a previous key frame adjacent to the first key frame in the video; and/or
wherein the fusion features of the second key frame are obtained by performing fusion on the features of the second key frame and fusion features of a third key frame in the video with a detection sequence preceding that of the second key frame.

3. The method according to claim 1 or 2, prior to the performing fusion on the features of the first key frame and fusion features of a second key frame in the video, the method further comprising:

obtaining an optical flow field between the first key frame and the second key frame; and

aligning the fusion features of the second key frame with the first key frame according to the optical flow field between the first key frame and the second key frame;

wherein the performing fusion on the features of the first key frame and fusion features of a second key frame in the video comprises: performing fusion on the aligned fusion features of the second key frame and the features of the first key frame.

4. The method according to any one of claims 1 to 3, the performing fusion on the features of the first key frame and fusion features of a second key frame in the video comprises:

determining a weight coefficient of the features of the first key frame and a weight coefficient of the fusion features of the second key frame; and

performing fusion on the features of the first key frame and the fusion features of the second key frame according to the weight coefficient of the features of the first key frame and the weight coefficient of the fusion features of the second key frame.

5. The method according to any one of claims 1 to 4, further comprising:

determining whether the first key frame is an initial key frame in the video;

wherein the performing fusion on the features of the first key frame and fusion features of a second key frame in the video comprises: in response to determining that the first key frame is not the initial key frame in the video, performing fusion on the features of the first key frame and the fusion features of the second key frame; and/or

in response to determining that the first key frame is the initial key frame in the video, obtaining an object detection result of the first key frame according to the features of the first key frame.

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

using an image frame randomly selected from the video as the first key frame; or

using an image frame selected from the video at an interval of several frames as the first key frame; or

splitting the video into a plurality of short video sequences, and separately selecting an image frame from each short video sequence as a key frame of each short video sequence, the key frames of the plurality of short video sequences comprising the first key frame; or

selecting a short video sequence from the video, and using an image frame selected from the selected short video sequence as the first key frame.

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

obtaining an optical flow field between a first non-key frame and the first key frame;

obtaining features of the first non-key frame according to the optical flow field between the first non-key frame and the first key frame, as well as the fusion features of the first key frame; and

performing detection on the first non-key frame according to the features of the first non-key frame to obtain an object detection result of the first non-key frame.

8. The method according to claim 7, wherein after the obtaining an optical flow field between a first non-key frame and the first key frame, the method further comprises:

aligning the fusion features of the first key frame with the first non-key frame according to the optical flow field between the first non-key frame and the first key frame;

and wherein the obtaining features of the first non-key frame according to the optical flow field between the first non-key frame and the first key frame, and the fusion features of the first key frame comprises:

obtaining the features of the first non-key frame according to the optical flow field between the first non-key frame and the first key frame, as well as the aligned fusion features of the first key frame.

9. The method according to any one of claims 1 to 8, wherein the image frame in the video has tag information;

and wherein the method further comprises:

training, based on a difference between an object detection result of at least one image frame in the video and the tag information, a video recognition network for implementing the method for video recognition, wherein the at least one image frame comprises the first key frame;

wherein the tag information comprises: position information and category information of at least one object contained in the image frame;

the object detection result comprises: an object detection box and an object category corresponding to the object detection box; and

the training, based on a difference between an object detection result of at least one image frame in the video and the tag information, a video recognition network for implementing the method for video recognition comprises: adjusting network parameter values in the video recognition network based on at least one of the following: based on the difference between the object detection box in the object detection result of the at least one image frame in the video and the position information in the tag information, or based on a difference between the object category

corresponding to the object detection box and the category information in the tag information.

10. A method for training a video recognition network, comprising:

extracting, using a video recognition network, features of a current key frame in a video sample and features of a previous key frame in the video sample;

performing, using the video recognition network, fusion on the features of the current key frame and the features of the previous key frame to obtain fusion features of the current key frame, and obtaining an optical flow field between a current target frame in the video sample and the current key frame, wherein the current target frame and the current key frame belong to a same segment of the video sample;

obtaining, using the video recognition network, features of the current target frame according to the optical flow field between the current target frame and the current key frame, and the fusion features of the current key frame;

performing, using the video recognition network, detection on the current target frame according to the features of the current target frame to obtain an object detection result of the current target frame; and

adjusting network parameter values of the video recognition network based on a difference between the object detection result of the current target frame and tag information of the current target frame.

11. The method according to claim 10, wherein the performing fusion on features of the current key frame and features of the previous key frame comprises:

obtaining an optical flow field between the current key frame and the previous key frame;

aligning the features of the previous key frame with the current key frame according to the optical flow field between the current key frame and the previous key frame; and

performing fusion on the aligned features of the previous key frame and the features of the current key frame.

12. The method according to claim 10 or 11, wherein the performing fusion on the aligned features of the previous key frame and the features of the current key frame comprises:

determining a weight coefficient of the features of the previous key frame and a weight coefficient of the features of the current key frame; and

performing fusion on the features of the previous key frame and the features of the current key frame according to the weight coefficient of the features of the previous key frame and the weight coefficient of the features of the current key frame.

13. The method according to any one of claims 10 to 12, wherein prior to the extracting the features of the current key frame in the video sample and the features of the previous key frame, the method further comprises:

using an image frame selected from the video sample as the current target frame; and

using an image frame in the video sample sampled forward a first distance by using the current target frame as a starting point, as the previous key frame, and using an image frame in the video sample sampled backward a second distance by using the current target frame as the starting point, as the current key frame.

14. The method according to claim 13, wherein the second distance is less than the first distance.

15. The method according to any one of claims 10 to 14, wherein the tag information of the current target frame comprises: position information and category information of at least one object contained in the current target frame;

the object detection result of the current target frame comprises: an object detection box in the current target frame and an object category corresponding to the object detection box;

the training the video recognition network based on the difference between the object detection result of the current target frame and the tag information of the current target frame comprises:

adjusting network parameter values in the video recognition network based on at least one of the following: based on a difference between the object detection box in

the object detection result of the current target frame and the position information in the tag information, or based on a difference between the object category corresponding to the object detection box and the category information in the tag information.

16. A method for recognizing a video, comprising:

obtaining a to-be-recognized video; and

recognizing, using a video recognition network, the video to obtain a video recognition result, wherein the video recognition network is obtained by training in the training method according to any one of claims 10 to 15.

17. An apparatus for recognizing a video, comprising:

a feature extraction network, configured to extract features of a first key frame in the video;

a fusion network, configured to perform fusion on the features of the first key frame and fusion features of a second key frame in the video to obtain fusion features of the first key frame, wherein a detection sequence of the second key frame in the video precedes that of the first key frame; and

a detection network, configured to perform detection on the first key frame according to the fusion features of the first key frame to obtain an object detection result of the first key frame.

18. An apparatus for training a video recognition network, comprising:

a video recognition network, configured to: extract features of a current key frame in a video sample and features of a previous key frame; perform fusion on the features of the current key frame and the features of the previous key frame to obtain fusion features of the current key frame, and obtain an optical flow field between a current target frame in the video sample and the current key frame, wherein the current target frame and the current key frame belong to a same segment of the video sample; obtain features of the current target frame according to the optical flow field between the current target frame and the current key frame, and the fusion features of the current key frame; and perform detection on the current target frame according to the features of the current target frame to obtain an object detection result of the current target frame; and

a training module, configured to adjust network parameter values of the video recognition network based on a difference between the object detection result of the current target frame and tag information of the current target frame.

19. An electronic device, comprising:

a processor; and

a memory for storing computer-readable instructions executable by the processor,

wherein execution of the computer-readable instructions by the processor causes the processor to implement the method according to any one of claims 1 to 16.

20. A computer storage medium, configured to store computer-readable instructions, wherein execution of the computer-readable instructions in a device causes a processor in the device to perform the method according to any one of claims 1 to 16.