

WHAT IS CLAIMED IS:

1. A method of detecting target object, comprising:
 - extracting, through a neural network, a feature of a reference frame and a feature
 - 5 of a frame under detection respectively, wherein
 - the frame under detection indicates a current frame in a video sequence in which a target object is to be detected or indicates an area image in the current frame which probably contains the target object,
 - the reference frame indicates a bounding box image of the target object in a
 - 10 frame for which a bounding box for the target object is determined, and a detection timing of the reference frame in the video sequence is before that of the frame under detection, and
 - an image size of the reference frame is smaller than that of the frame under detection;
 - 15 inputting each of at least two feature groups output from at least two network layers of the neural network into a detector so to obtain a corresponding detection result group output from the detector, wherein
 - each of the feature groups comprises the feature of the reference frame and the feature of the frame under detection, and
 - 20 each of the detection result groups comprises a classification result and a regression result with respect to a plurality of candidate boxes for the target object; and
 - determining a bounding box for the target object in the frame under detection according to the at least two detection result groups output from the detector for the at

least two feature groups.

2. The method according to claim 1, further comprising:

5 taking the bounding box for the target object in the frame under detection as a next reference frame;

extracting, through the neural network, a feature of the next reference frame and a feature of a next frame under detection respectively, wherein a timing of the next frame under detection in the video sequence is after that of the frame under detection;

10 inputting each of at least two feature groups output from at least two network layers of the neural network into a detector so as to obtain a corresponding detection result group output from the detector, wherein each of the feature groups comprises a feature of the next reference frame and a feature of the next frame under detection; and

15 determining, according to the at least two detection result groups output from the detector for the at least two feature groups, a bounding box for the target object in the next frame under detection.

3. The method according to claim 1 or 2, wherein

20 the neural network comprises a plurality of convolution layers; inputting each of at least two feature groups output from at least two network layers of the neural network into the detector, comprises: inputting at least two feature groups output from at least two convolution layers of the neural network into the detector respectively; or

the neural network comprises a plurality of convolution blocks, and each

convolution block comprises at least one convolution layer,

inputting each of at least two feature groups output from at least two network layers of the neural network into the detector, comprises: inputting at least two feature groups output from at least two convolution blocks of the neural network
5 into the detector respectively; or

the neural network comprises at least one convolution block and at least one convolution layer, and each of the at least one convolution block comprises at least one convolution layer;

inputting each of at least two feature groups output from at least two
10 network layers of the neural network into the detector, comprises: inputting at least two feature groups output from at least one convolution block of the neural network and from at least one convolution layer of the neural network into the detector.

4. The method according to any one of claims 1-3, wherein in a case that a frame
15 under detection indicates an area image in the current frame which probably contains the target object, and the current frame is to be subjected to target object detection, the method further comprises:

cutting the area image from the current frame as the frame under detection, wherein, at least one of a length and a width of the area image is greater than that of
20 the reference frame, and the area image is centered at a center point of the reference frame.

5. The method according to any one of claims 1-4, wherein inputting each of the at least two feature groups output from the at least two network layers of the neural

network into the detector so as to obtain a corresponding detection result group output from the detector, comprises:

acquiring, based on the feature of the reference frame, a classification weight and a regression weight for the detector; and

5 processing the feature of the frame under detection with the classification weight and the regression weight, respectively, so as to obtain a classification result and a regression result with respect to each of the plurality of candidate boxes output from the detector.

10 6. The method according to claim 5, wherein,

acquiring, based on the feature of the reference frame, the classification weight for the detector comprises:

performing, through a first convolution layer of the neural network, convolution operation on the feature of the reference frame, and taking a first feature

15 obtained through the convolution operation as the classification weight for the detector;

acquiring, based on the feature of the reference frame, the regression weight for the detector comprises:

performing, through a second convolution layer of the neural network, convolution operation on the feature of the reference frame, and taking a second

20 feature obtained through the convolution operation as the regression weight for the detector;

processing the feature of the frame under detection with the classification weight and the regression weight, respectively, so as to obtain the classification result and the

regression result with respect to each of the plurality of candidate boxes output from the detector comprises:

performing, with the classification weight, convolution operation on the feature of the frame under detection so as to obtain a classification result with respect to each of the plurality of candidate boxes; and

performing, with the regression weight, convolution operation on the feature of the frame under detection so as to obtain a regression result with respect to each of the plurality of candidate boxes.

7. The method according to claim 6, wherein,

performing convolution operation, with the classification weight, on the feature of the frame under detection so as to obtain the classification result with respect to each of the plurality of candidate boxes comprises:

performing, through a third convolution layer of the detector, convolution operation on the feature of the frame under detection so as to obtain a third feature, the number of channels of which is identical to the number of channels of the feature of the frame under detection; and

performing, with the classification weight, convolution operation on the third feature so as to obtain the classification result with respect to each of the plurality of candidate boxes;

performing, with the regression weight, convolution operation on the feature of the frame under detection so as to obtain a regression result with respect to each of the plurality of candidate boxes comprises:

performing, through a fourth convolution layer of the detector, convolution

operation on the feature of the frame under detection so as to obtain a fourth feature, the number of channels of which is identical to the number of channels of the feature of the frame under detection; and

performing, with the regression weight, on the fourth feature so as to obtain
5 the regression result with respect to each of the plurality of candidate boxes.

8. The method according to claim 5, wherein,

acquiring, based on the feature of the reference frame, the classification weight for the detector comprises:

10 performing, through a fifth convolution layer of the detector, convolution operation on the feature of the reference frame, and taking L fifth features obtained through the convolution operation as the classification weight for the detector, wherein the number of channels of the feature of the reference frame is M1, the number of channels of the fifth feature is N1, wherein each of M1, N1, and L is an
15 integer greater than 0, and M1 is greater than N1;

processing the feature of the frame under detection with the classification weight so as to obtain the classification results with respect to each of the plurality of candidate boxes output from the detector comprises:

performing, through a sixth convolution layer of the detector, convolution
20 operation on the feature of the frame under detection so as to obtain L sixth features with L channels, wherein the number of channels of the feature of the frame under detection is M2, which is an integer greater than 0 and is greater than N1;

performing, with each of the L fifth features of the classification weight, convolution operation on a corresponding feature of the L sixth features, respectively,

so as to obtain L convolution results; and

performing, through a seventh convolution layer of the detector, convolution operation on the L convolution results so as to obtain the classification result with respect to each of the plurality of candidate boxes.

5

9. The method according to claim 8, wherein after obtaining the L convolution results, the method further comprises connecting the L convolution results; and

performing, through the seventh convolution layer of the detector, convolution operation on the L convolution results comprises:

10 performing, through the seventh convolution layer of the detector, convolution operation on the connected L convolution results.

10. The method according to claim 5, wherein,

15 acquiring the regression weight for the detector based on the feature of the reference frame comprises:

performing, through an eighth convolution layer of the detector, convolution operation on the feature of the reference frame, and taking eighth features with P channels obtained through the convolution operation as the regression weight for the detector; wherein the number of channels of the feature of the reference frame is M1, 20 the number of channels of the eighth feature is N2, each of M1, N2, and P is an integer greater than 0, and M1 is greater than N2;

processing the feature of the frame under detection with the regression weight, so as to obtain the regression result with respect to each of the plurality of candidate boxes output from the detector comprises:

performing, through a ninth convolution layer of the detector, convolution operation on the feature of the frame under detection, so as to obtain P ninth features with N2 channels, wherein the number of channels of the feature of the frame under detection is M2, which is an integer greater than 0 and is greater than N2;

5 performing, with each of the P eighth features of the regression weight, convolution operation on a corresponding feature of the P ninth features, so as to obtain P convolution results; and

performing, through a tenth convolution layer of the detector, convolution operation on the P convolution results, so as to obtain the regression result with
10 respect to each of the plurality of candidate boxes.

11. The method according to claim 10, wherein, after obtaining the P convolution results, the method further comprises connecting the P convolution results, performing, through the tenth convolution layer of the detector, convolution operation
15 on the P convolution results comprises:

performing, through the tenth convolution layer of the detector, convolution operation on the connected P convolution results.

12. The method according to any one of claims 1 to 11, wherein determining the
20 bounding box for the target object in the frame under detection according to the at least two detection result groups output from the detector comprises:

performing one of weighting, averaging, and weighted-averaging on the classification results with respect to the at least two detection result groups so as to obtain a comprehensive classification result with respect to the plurality of candidate

boxes; and/or performing one of weighting, averaging, and weighted-averaging on the regression results with respect to the at least two detection result groups, so as to obtain a comprehensive regression result with respect to the plurality of candidate boxes; and

- 5 acquiring a bounding box for the target object in the frame under detection according to the comprehensive classification result and the comprehensive regression result with respect to the plurality of candidate boxes.

13. The method according to claim 12, wherein acquiring the bounding box for
10 the target object in the frame under detection according to the comprehensive classification result and the comprehensive regression result with respect to the plurality of candidate boxes comprises:

- selecting a candidate box from the plurality of candidate boxes according to the comprehensive classification result and the comprehensive regression result with
15 respect to the plurality of candidate boxes, and performing regression on the selected candidate box according to an offset of the comprehensive regression result with respect to the selected candidate box so as to obtain the bounding box for the target object in the frame under detection.

- 20 14. The method according to claim 13, wherein,

 selecting the candidate box from the plurality of candidate boxes according to the comprehensive classification result and the comprehensive regression result with respect to the plurality of candidate boxes comprises:

 selecting the candidate box from the plurality of candidate boxes according

to a weight coefficient for the comprehensive classification result and a weight coefficient for the comprehensive regression result.

after obtaining the comprehensive regression result with respect to the plurality of candidate boxes, the method further comprises:

- 5 adjusting the comprehensive classification result according to the comprehensive regression result with respect to the plurality of candidate boxes;
 selecting the candidate box from the plurality of candidate boxes according to the comprehensive classification result and the comprehensive regression result with respect to the plurality of candidate boxes comprises: selecting the candidate box
- 10 from the plurality of candidate boxes according to the adjusted comprehensive classification result.

15. The method according to any one of claims 1 to 14, further comprising at least one of the following:

- 15 labeling, after acquiring the bounding box for the target object in the frame under detection, the bounding box for the target object in the video sequence;
 determining, after acquiring bounding boxes for the target object in a plurality of frames in the video sequence, a movement trajectory of the target object in the video sequence according to the positions of the bounding boxes for the target object in the
- 20 plurality of frames in the video sequence;
 adjusting, after acquiring the bounding boxes for the target object in the plurality of frames in the video sequence, shooting parameters of a camera according to the position of the bounding box for the target object;
 determining, after acquiring the bounding box for the target object in the frame

under detection, whether a position of the bounding box for the target object in the frame under detection is within a preset area; and output a prompt message in response to a position of the bounding box for the target object within the preset area;

5 matching, in a case that there are a plurality of target objects, after acquiring the bounding boxes for the plurality of target objects in the frame under detection, the plurality of target objects in the frame under detection with the bounding boxes for the plurality of target objects in the reference frame, and labeling identification for each of the plurality of target objects in the frame under detection according to a matching result and the identification of each of the plurality of target object in the reference
10 frame;

 determining, in a case that there are a plurality of target objects, after acquiring the bounding boxes for the plurality of target objects in multiple frames in the video sequence, movement trajectory of the plurality of target objects in the video sequence according to positions of the bounding boxes for each of the plurality of target objects
15 in the multiple frames in the video sequence; and tracking the plurality of target objects according to the movement trajectory of the plurality of target objects in the video sequence;

 analyzing, in a case that there are a plurality of target objects, after acquiring the bounding boxes for the plurality of target objects in multiple frames in the video
20 sequence, behaviors of the plurality of target objects according to changes in the number and/or position of the plurality of target objects in multiple frames in the video sequence; and

 determining, after acquiring the bounding boxes for the target object in multiple frames in the video sequence, motion state of the target object according to positions

of the bounding box for the target object in the multiple frames in the video sequence, and, performing intelligent driving control on a carrier on which a camera shooting the video sequence is carried based on the motion state of the target object and motion state of the carrier.

5

16. A method of training target object detection network, comprising:

inputting at least one sample pair into a neural network, wherein each of the at least one sample pair comprises a reference frame and a frame under detection which correspond to each other, and outputting a bounding box for a target object in the
10 frame under detection by using the method of detecting target object according to any one of claims 1 to 15; and

taking the output bounding box for the target object in the frame under detection as a predicted bounding box, and training the neural network and the detector according to label information in the frame under detection and the predicted
15 bounding box.

17. A device for detecting target object, comprising:

a neural network, configured to extract a feature of a reference frame and a feature of a frame under detection, respectively; wherein the frame under detection
20 indicates a current frame in a video sequence in which a target object is to be detected or indicates an area image in the current frame which probably contains a target object, the reference frame indicates a bounding box image of the target object in a frame for which a bounding box for the target object is determined, and a detection timing of the reference frame in the video sequence is before that of the frame under detection,

and an image size of the reference frame is smaller than that of the frame under detection;

5 a detector, configured to receive each of at least two feature groups output from at least two network layers of the neural network, and to output a detection result group for each feature group; wherein each feature group comprise the feature of the reference frame and the feature of the frame under detection, and each detection result group comprises a classification result and a regression result with respect to each of a plurality of candidate boxes for the target object; and

10 an acquiring module, configured to determine a bounding box for the target object in the frame under detection according to at least two detection result groups output from the detector for the at least two feature groups.

18. A device for training target object detection network, comprising:

15 a target object detection network, configured to receive at least one sample pair, each of the at least one sample pair including a reference frame and a frame under detection which correspond to each other, and to output a bounding box for the target object in the frame under detection of the sample pair, wherein the target object detection network comprises the device for detecting target object according to claim 17; and

20 a training module, configured to take the bounding box for the target object in the frame under detection as a predicted bounding box, and to train the neural network and the detector according to label information of the frame under detection and the predicted bounding box.

19. An electronic apparatus, comprising:
memory, configured to store a computer program; and
a processor, configured to execute the computer program stored in the memory,
and to implement the method according to any one of claims 1 to 16 in a case that the
5 computer program is executed.

20. A computer readable storage medium, having a computer program stored
thereon, wherein in a case that the computer program is executed by a processor, the
method according to any one of claims 1 to 16 is implemented.

10