

CLAIMS

What is claimed is:

[Claim 1] A video-based human action recognition method, comprising:

obtaining a plurality of multi-human image frames in a video, wherein each multi-human
5 image frame comprises a plurality of human instances;

generating detected human bounding boxes of human instances in each multi-human
image frame;

determining an image scene of each multi-human image frame, one image scene
corresponding to one or more categories of actions, each image scene is coupled to the
10 corresponding action, and each image scene is configured to reflect the action of the human
instances; and

recognizing the actions of human instances in the plurality of multi-human image frames
based on the detected human bounding boxes and the image scenes of the plurality of multi-
human image frames.

15 [Claim 2] The method according to claim 1, wherein the recognizing the actions of human
instances in the plurality of multi-human image frames based on the detected human bounding
boxes and the image scenes of the plurality of multi-human image frames comprises:

recognizing the actions of the human instances in the plurality of multi-human image
frames based on the detected human bounding boxes of the plurality of multi-human image
20 frames, to obtain an initial probability of each of a plurality of predefined actions corresponding
to the human instance;

correcting the initial probability of each of the predefined multi-category actions
corresponding to the human instance based on the image scenes of the plurality of multi-human
image frames, to obtain a corrected probability of each predefined action; and

25 taking the predefined action with the highest probability in each multi-human image frame
as the action recognized by the human instance.

[Claim 3] The method according to claim 2, wherein the correcting the initial probability
of each of the predefined multi-category actions corresponding to the human instance based on

the image scenes of the plurality of multi-human image frames, to obtain a corrected probability of each predefined action comprises:

correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames and the image scenes of the plurality of multi-human image frames, to obtain the corrected probability of each predefined action.

[Claim 4] The method according to claim 3, wherein the degree of overlap between adjacent detected human bounding boxes is represented by an Intersection over Union (IoU) between adjacent detected human bounding boxes.

[Claim 5] The method according to claim 3, wherein the correcting the initial probability of each of the predefined actions corresponding to the human instance based on the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames comprises:

when the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames is greater than a predetermined threshold, increasing a probability of a predefined action related to a plurality of persons among the plurality of predefined actions corresponding to the human instances; or

when the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames is less than or equal to the predetermined threshold, increasing a probability of a predefined action related to a single person among the plurality of predefined actions corresponding to the human instances.

[Claim 6] The method according to any one of claims 2 to 5, wherein the correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the image scenes of the plurality of multi-human image frames comprises:

for the human instance in a region where the image scenes are located in the plurality of multi-human image frames, increasing probabilities of the predefined multi-category actions corresponding to the human instance and the predefined action corresponding to the determined

image scene based on the image scenes of the plurality of multi-human image frames.

[Claim 7] The method according to claim 1, wherein the generating detected human bounding boxes of human instances in each multi-human image frame comprises:

5 generating initial human bounding boxes of human instances in each multi-human image frame;

correcting an initial human bounding box of the multi-human image based on position difference data between the multi-human image and the previous image frame of the multi-human image, to obtain the detected human bounding box, wherein the position difference data comprises a position difference of the human instance in the multi-human image between the
10 previous image frame and the multi-human image; and

recognizing a pose of the human instance in the multi-human image based on the detected human bounding box of the multi-human image.

[Claim 8] The method according to claim 7, wherein the correcting an initial human bounding box of the multi-human image based on position difference data between the multi-
15 human image and the previous image frame of the multi-human image, to obtain the detected human bounding box comprises:

correcting the i-th initial human bounding box in the t-th image frame in the plurality of multi-human image frames by using the following formula to obtain the i-th detected human bounding box:

20
$$\hat{I}_i^t = I_i^{t-1} + F^{t-1 \rightarrow t},$$

$$\tilde{I}_i^k = \alpha \hat{I}_i^t + (1 - \alpha) I_i^t,$$

wherein, I_i^{t-1} represents a coordinate set of the i-th initial human bounding box in the (t-1)-th image frame in the plurality of multi-human image frames, $F^{t-1 \rightarrow t}$ represents the position difference data between the (t-1)-th image frame and the t-th image frame, $\hat{I}_i^t$ represents a
25 coordinate set of the i-th predicted human bounding box in the t-th image frame obtained based on $F^{t-1 \rightarrow t}$, I_i^t represents a coordinate set of the i-th initial human bounding box in the t-th image frame, $\tilde{I}_i^k$ represents a coordinate set of the i-th detected human bounding box in the t-th image frame, and α represents a weight coefficient synthesized by I_i^t and $\hat{I}_i^t$, wherein i and

t are positive integers.

[Claim 9] The method according to claim 1, wherein the initial human bounding boxes of the human instances in each multi-human image frame are generated by a first neural network model, and a model network of the first neural network model comprises at least one of the following: ResNet152 network, ResNeXt101 network, or SeNet154 network.

[Claim 10] A video-based human action recognition apparatus, comprising:
an acquisition module, configured to obtain a plurality of multi-human image frames in a video, wherein each multi-human image frame comprises a plurality of human instances;
a first generation module, configured to generate detected human bounding boxes of human instances in each multi-human image frame;
a determining module, configured to determine an image scene of each multi-human image frame, one image scene corresponding to one or more categories of actions, each image scene is coupled to the corresponding action, and each image scene is configured to reflect the action of the human instances; and
a recognition module, configured to recognize the actions of human instances in the plurality of multi-human image frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames.

[Claim 11] A readable medium, wherein the readable medium stores instructions, and when the instructions are executed on an electronic device, the electronic device is enabled to perform the video-based human action recognition method according to any one of claims 1 to 9.

[Claim 12] An electronic device, comprising:
a memory, configured to store instructions to be executed by one or more processors of the electronic device; and
a processor, one of the processors of the electronic device, configured to perform the video-based human action recognition method according to any one of claims 1 to 9.