

CLAIMS

What is claimed is:

[Claim 1] A video-based human pose estimation 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 initial human bounding boxes of human instances in each multi-human image
frame, each human bounding box encloses a human instance, and configured to mark the
position of the enclosed human instance in the corresponding image;

correcting an initial human bounding box of the multi-human image based on position
10 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
previous image frame and the multi-human image, the position difference data between two
adjacent multi-human images is the optical flow between the two adjacent multi-human images;
15 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 2] The method according to claim 1, wherein the correcting an initial human
bounding box of the multi-human image based on position difference data between the multi-
20 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 formulas to obtain the i-th detected human
bounding box:

$$\begin{aligned} \hat{I}_i^t &= I_i^{t-1} + F^{t-1 \rightarrow t}, \\ \tilde{I}_i^k &= \alpha \hat{I}_i^t + (1 - \alpha) I_i^t, \end{aligned}$$

wherein I_i^{t-1} represents a coordinate point 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 coordinate point 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 point set of the i-th initial human bounding box in the t-th image frame, $\tilde{I}_i^k$ represents a coordinate point 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 $\tilde{I}_i^t$, wherein i and t are positive integers.

[Claim 3] The method according to claim 1 or 2, wherein the recognizing a pose of the human instance in the multi-human image based on the detected human bounding box of the multi-human image comprises:

generating initial pose key points of the human instances in each multi-human image frame based on the detected human bounding boxes of the multi-human image;

correcting the initial pose key points of the human instances in the multi-human image based on the position difference data between the multi-human image and the previous image frame of the multi-human image and position difference data between the multi-human image and the next image frame of the multi-human image, to obtain detected pose key points, wherein the position difference data between the multi-human image and the next image frame comprises a position difference of the human instances in the multi-human image between the multi-human image and the next image frame; and

recognizing poses of the human instances in the multi-human image based on the detected pose key points in the multi-human image.

[Claim 4] The method according to claim 3, wherein the correcting the initial pose key points of the human instances in the multi-human image based on the position difference data between the multi-human image and the previous image frame of the multi-human image and position difference data between the multi-human image and the next image frame of the multi-human image, to obtain detected pose key points comprises:

correcting initial pose key points of the human instances in the i-th detected human

bounding box in the t-th image frame in the plurality of multi-human image frames by using the following formulas to obtain detected pose key points:

$$\hat{J}_i^t = J_i^{t-1} + F^{t-1 \rightarrow t},$$

$$\hat{J}_i^t = J_i^{t+1} - F^{t \rightarrow t+1},$$

$$\tilde{J}_i^t = \beta \cdot \hat{J}_i^t + \beta \cdot \hat{J}_i^t + (1 - 2\beta) \cdot J_i^t,$$

wherein J_i^{t-1} represents a coordinate point set of initial pose key points of human instances in the i-th detected human bounding box in the (t-1)-th image frame in the plurality of multi-human image frames, $F^{t-1 \rightarrow t}$ represents position difference data between the (t-1)-th image frame and the t-th image frame, $\hat{J}_i^t$ represents a coordinate point set of predicted pose key points of human instances in the i-th detected human bounding box in the t-th image frame obtained based on $F^{t-1 \rightarrow t}$; J_i^{t+1} represents a coordinate point set of initial pose key points of human instances in the i-th detected human bounding box in the (t+1)-th image frame in the plurality of multi-human image frames; $F^{t \rightarrow t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\hat{J}_i^t$ represents a coordinate point set of predicted pose key points of human instances in the i-th detected human bounding box in the t-th image frame obtained based on $F^{t \rightarrow t+1}$; $\tilde{J}_i^t$ represents a coordinate point set of detected pose key points of human instances in the i-th detected human bounding box in the t-th image frame, and β represents a weight coefficient synthesized between J_i^t , $\hat{J}_i^t$ and $\hat{J}_i^t$, wherein i and t are positive integers.

[Claim 5] The method according to claim 1, wherein before the recognizing a pose of the human instance in the multi-human image based on the detected human bounding box of the multi-human image, the method further comprises:

generating identification information corresponding to the detected human bounding boxes of the human instances in each multi-human image frame, wherein one piece of identification information is used for identifying the detected human bounding box of one

human instance.

[Claim 6] The method according to claim 5, wherein the recognizing a pose of the human instance in the multi-human image based on the detected human bounding box of the multi-human image comprises:

5 recognizing poses of the human instances in the multi-human image based on the detected human bounding boxes of the multi-human image and the identification information of the detected human bounding boxes.

[Claim 7] The method according to claim 5 or 6, wherein the generating identification information corresponding to the detected human bounding boxes of the human instances in
10 each multi-human image frame comprises:

 generating the identification information corresponding to the detected human bounding boxes of the human instances in each multi-human image frame based on appearance features of the human instances in each multi-human image frame and positions of the human instances in the image.

15 [Claim 8] The method according to claim 5, further comprising:

 tracking the human instances in the plurality of multi-human image frames based on the identification information corresponding to the detected human bounding boxes of the human instances in each multi-human image frame and the poses of the human instances.

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

[Claim 10] The method according to claim 3, wherein the initial pose key points of the human instances in each multi-human image frame are generated by a second neural network
25 model, and a backbone network of the second neural network model comprises at least one of the following: HRNet and SimpleNet.

[Claim 11] A video-based human pose estimation method, comprising:

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

generating detected human bounding boxes of human instances in each multi-human image frame, each human bounding box encloses a human instance, and configured to mark the position of the enclosed human instance in the corresponding image;

generating initial pose key points of the human instances in each multi-human image frame based on the detected human bounding boxes of the multi-human image;

correcting the initial pose key points of the human instances in the multi-human image based on the position difference data between the multi-human image and the previous image frame of the multi-human image and position difference data between the multi-human image and the next image frame of the multi-human image, to obtain detected pose key points, wherein the position difference data between the multi-human image and the previous image frame comprises a position difference of the human instances in the multi-human image between the multi-human image and the previous image frame, and the position difference data between the multi-human image and the next image frame comprises a position difference of the human instances in the multi-human image between the multi-human image and the next image frame, the position difference data between two adjacent multi-human images is the optical flow between the two adjacent multi-human images; and

recognizing poses of the human instances in the multi-human image based on the detected pose key points in the multi-human image.

[Claim 12] The method according to claim 11, wherein the correcting the initial pose key points of the human instances in the multi-human image based on the position difference data between the multi-human image and the previous image frame of the multi-human image and position difference data between the multi-human image and the next image frame of the multi-human image, to obtain detected pose key points comprises:

correcting initial pose key points of the human instances in the i-th detected human bounding box in the t-th image frame in the plurality of multi-human image frames by using

the following formulas to obtain detected pose key points:

$$\hat{J}_i^t = J_i^{t-1} + F^{t-1 \rightarrow t},$$

$$\hat{J}_i^{t+1} = J_i^t + F^{t \rightarrow t+1},$$

$$\tilde{J}_i^t = \beta \cdot \hat{J}_i^t + \beta \cdot \hat{J}_i^{t+1} + (1 - 2\beta) \cdot J_i^t,$$

5 wherein J_i^{t-1} represents a coordinate point set of initial pose key points of human instances in the i-th detected human bounding box in the (t-1)-th image frame in the plurality of multi-human image frames, $F^{t-1 \rightarrow t}$ represents position difference data between the (t-1)-th image frame and the t-th image frame, $\hat{J}_i^t$ represents a coordinate point set of predicted pose key points of human instances in the i-th detected human bounding box in the t-th image frame
 10 obtained based on $F^{t-1 \rightarrow t}$; J_i^{t+1} represents a coordinate point set of initial pose key points of human instances in the i-th detected human bounding box in the (t+1)-th image frame in the plurality of multi-human image frames; $F^{t \rightarrow t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\hat{J}_i^{t+1}$ represents a coordinate point set of predicted pose key points of human instances in the i-th detected human bounding box in the t-th image frame obtained based on $F^{t \rightarrow t+1}$; $\tilde{J}_i^t$ represents a coordinate point set of detected pose
 15 key points of human instances in the i-th detected human bounding box in the t-th image frame, and β represents a weight coefficient synthesized between J_i^t , $\hat{J}_i^t$ and $\hat{J}_i^{t+1}$, wherein i and t are positive integers.

[Claim 13] The method according to claim 11 or 12, wherein the generating detected
 20 human bounding boxes of human instances in each multi-human image frame comprises:

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-
 25 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 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.

5 [Claim 14] The method according to claim 13, wherein the 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 comprises:

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

$$\begin{aligned}\hat{I}_i^t &= I_i^{t-1} + F^{t-1 \rightarrow t}, \\ \tilde{I}_i^k &= \alpha \hat{I}_i^t + (1 - \alpha) I_i^t,\end{aligned}$$

wherein I_i^{t-1} represents a coordinate point set of the i-th initial human bounding box in
15 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 coordinate point 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 point set of the i-th initial human bounding box in the t-th image frame, $\tilde{I}_i^k$ represents a coordinate point set of the i-th detected
20 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 15] A video-based human pose estimation apparatus, comprising:

an image 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
25 instances;

a human detection module, configured to generate initial human bounding boxes of the human instances in each multi-human image frame, each human bounding box encloses a

human instance, and configured to mark the position of the enclosed human instance in the corresponding image;

5 a human correction module, configured to correct 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 previous image frame and the multi-human image, the position difference data between two adjacent multi-human images is the optical flow between the two adjacent multi-human images; and

10 a pose estimation module, configured to recognize a pose of the human instance in the multi-human image based on the detected human bounding box of the multi-human image.

[Claim 16] A video-based human pose estimation apparatus, comprising:

15 an image 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, each human bounding box encloses a human instance, and configured to mark the position of the enclosed human instance in the corresponding image;

a human detection module, configured to generate detected human bounding boxes of the human instances in each multi-human image frame;

20 a pose detection module, configured to generate initial pose key points of the human instances in each multi-human image frame based on the detected human bounding boxes of the multi-human image;

25 a pose correction module, configured to correct the initial pose key points of the human instances in the multi-human image based on the position difference data between the multi-human image and the previous image frame of the multi-human image and position difference data between the multi-human image and the next image frame of the multi-human image, to obtain detected pose key points, wherein the position difference data between the multi-human image and the previous image frame comprises a position difference of the human instances in

the multi-human image between the multi-human image and the previous image frame, and the position difference data between the multi-human image and the next image frame comprises a position difference of the human instances in the multi-human image between the multi-human image and the next image frame, the position difference data between two adjacent multi-human images is the optical flow between the two adjacent multi-human images; and

5 a pose estimation module, configured to recognize poses of the human instances in the multi-human image based on the detected pose key points in the multi-human image.

[Claim 17] 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

10 perform the video-based human pose estimation method according to any one of claims 1 to 14.

[Claim 18] 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-

15 based human pose estimation method according to any one of claims 1 to 14.