# TITLE OF INVENTION: VIDEO-BASED POSE ESTIMATION METHOD AND APPARATUS, MEDIUM, AND ELECTRONIC DEVICE

## TECHNICAL FIELD

**[0001]** The present application relates to the field of computer vision, and particularly discloses a video-based human pose estimation method and apparatus, a medium, and an electronic device.

## BACKGROUND

**[0002]** With the rapid development of artificial intelligence technologies such as machine vision and deep learning, the human pose estimation technology has also developed rapidly. Human pose estimation is of great significance in application fields such as activity recognition and motion capture. How to perform human pose estimation quickly and accurately has become a widespread concern.

**[0003]** In the prior art, the human pose estimation method mainly includes recognizing a human body and generating a bounding box through a human detection model, then obtaining human joint points through a pose estimation model, and then analyzing the joint points to obtain human poses. However, this method depends on the accuracy of output results of the human detection model and the pose estimation model. If the bounding box or joint point prediction is inaccurate, the poses estimated by using the joint points subsequently will be inaccurate.

## SUMMARY

**[0004]** Embodiments of the present application provide a video-based human pose

estimation method and apparatus, a medium, and an electronic device, which can correct detected human bounding boxes and pose key points of human instances in a plurality of multi-human image frames through optical flow between adjacent multi-human images, thereby improving the accuracy of the detected human bounding boxes and pose key points of the human instances in the plurality of multi-human image frames and improving the accuracy of human pose estimation of the human instances.

**[0005]** In a first aspect, the embodiments of the present application disclose a video-based human pose estimation method, the method including:

obtaining a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances; 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, where the position difference data includes 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.

**[0006]** It can be understood that a video containing a plurality of multi-human image frames is a video in a crowded scene, and the plurality of multi-human image frames may be all images obtained by performing framing processing on the video. The initial human bounding box of the above-mentioned human instance can be detected by a pre-trained human detection model or a conventional human detection model, where the pre-trained human detection model can be obtained through training based on different training data sets in the crowded scene, and supports the human instance detection in the crowded scene. In addition, the position difference data between two adjacent images may be an optical flow between the two images. For example, the position difference data between the multi-human image and the previous image frame of the multi-human image is the optical flow between the two images, so the position difference

data includes the position difference of the human instance in the multi-human image between the previous image frame and the multi-human image, which reflects the motion information of the human instance between the previous image frame and the multi-human image in the time domain, thereby further reflecting the position change in the human bounding box of the human instance from the previous image frame to the multi-human image. Clearly, because this solution corrects the detected human bounding box of each human instance by using the position difference data (i.e., the optical flow) between adjacent multi-human images, and the position difference data reflects the information about the time domain, with reference to the information about the time domain, missing or dropping frames in the process of detecting the human bounding boxes of the plurality of multi-human image frames can be avoided, thus implementing the smoothing processing on the human bounding boxes of the human instances in the plurality of multi-human image frames (i.e., the optical flow smoothing below). Furthermore, the human bounding boxes of the human instances in the plurality of multi-human image frames are corrected by using the position difference data between the adjacent multi-human images, that is, optical flow smoothing is performed on the human bounding boxes of the human instances in the plurality of multi-human image frames, which improves the accuracy of the human bounding boxes detected for the human instances in the video of the plurality of multi-human image frames, thereby helping improve the accuracy of subsequent human pose estimation based on the human bounding box.

**[0007]** In addition, in some embodiments, the initial human bounding box of the human instance in a multi-human image can also be corrected by using the position difference data between a next image frame of a multi-human image frame and the multi-human image frame, which is not limited in the embodiments of the present application.

**[0008]** In a possible implementation of the first aspect, 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 includes: 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:

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

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

where $I_i^{t-1}$ 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, $F^{t-1\to t}$ 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 $F^{t-1\to t}$, $I_i^t$ 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 $\alpha$ represents a weight coefficient synthesized by $I_i^t$ and $\hat{I}_i^t$, where i and t are positive integers.

**[0009]** It can be understood that a set of coordinate points of the bounding box of a human instance is used to represent the human bounding box. In addition, the $\hat{I}_i^t$ specifically refers to obtaining the i-th predicted human bounding box of the same human instance in the t-th image frame according to the motion status of a human instance in the (t–1)-th image frame reflected by $F^{t-1\to t}$ from the (t–1)-th image frame to the t-th image frame based on the i-th initial human bounding box in the (t–1)-th image frame, i.e., the position change in the i-th initial human bounding box in the (t–1)-th image frame from the (t–1)-th image frame to the t-th image frame. Furthermore, by synthesizing the coordinate point set (i.e., $\hat{I}_i^t$) of the i-th predicted human bounding box in the t-th image frame and the coordinate point set (i.e., $I_i^t$) of the i-th initial human bounding box in the t-th frame image, i.e., performing weighted average processing on the corresponding coordinates in two coordinate point sets, the i-th initial human bounding box in the t-th image frame can be corrected to obtain a more accurate i-th detected human bounding box.

**[0010]** In a possible implementation of the first aspect, 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 includes: 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, where the position difference data between the multi-human image and the next image frame includes 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.

**[0011]** It can be understood that the pose key points of human instances are the following human joint points (or joint points), which have different names but the same essence and are used to mark key points of human poses. The initial pose key points of the foregoing human instances can be detected by a pre-trained human pose estimation model or a conventional human pose estimation model, where the pre-trained human pose estimation model can be obtained through training based on a plurality of training data sets in crowded scenes, and supports detection of pose key points of human instances in crowded scenes, thereby implementing human pose estimation in crowded scenes. Specifically, the position difference data between the multi-human image and the previous image frame of the multi-human image is the optical flow between the two images, so the position difference data includes the position difference of the human instance in the multi-human image between the previous image frame and the multi-human image, which reflects the motion information of the human instance between the previous image frame and the multi-human image in the time domain, thereby further reflecting the position change in the pose key points of the human instance from the previous image frame to the multi-human image. Similarly, the position difference data between the foregoing multi-human image and the next image frame of the multi-human image is the optical flow between the two images, and reflects the position change in the pose key points of

the human instances from the multi-human image to the next image frame. Clearly, because this solution corrects the detected human bounding box of each human instance by using the position difference data (i.e., the optical flow) between adjacent multi-human images, and the position difference data reflects the information about the time domain, with reference to the information about the time domain, missing or dropping frames in the process of detecting pose key points of human instances in the plurality of multi-human image frames is avoided, thus implementing the smoothing processing on the human key points of the human instances in the plurality of multi-human image frames (i.e., the optical flow smoothing below). In this way, while the human bounding boxes detected from the human instances in a plurality of multi-human image frames are corrected based on the position difference data between adjacent multi-human images, the pose key points detected from the human instances in the plurality of multi-human image frames are corrected. That is, optical flow smoothing is performed on the pose key points of the human instances in the plurality of multi-human image frames, thereby further improving the accuracy of human pose estimation of human instances in a video containing the plurality of multi-human image frames.

**[0012]** In a possible implementation of the first aspect, 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 includes: 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:

$$\widehat{J'}_i^{\mathrm{t}}=J_i^{\mathrm{t}-1}+F^{\mathrm{t}-1\rightarrow\mathrm{t}},$$

$$\widehat{J''}_i^{\mathrm{t}}=J_i^{\mathrm{t}+1}-F^{\mathrm{t}\rightarrow\mathrm{t}+1},$$

$$\tilde{J}_i^{\mathrm{t}}=\beta\cdot\widehat{J'}_i^{\mathrm{t}}+\beta\cdot\widehat{J''}_i^{\mathrm{t}}+(1-2\beta)\cdot J_i^{\mathrm{t}},$$

where $J_i^{t-1}$ 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, $F^{t-1\to t}$ represents position difference data between the (t–1)-th image frame and the t-th image frame, $\widehat{J'}_i^t$ 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 $F^{t-1\to t}$; $J_i^{t+1}$ 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; $F^{t\to t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\widehat{J''}_i^t$ 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 $F^{t\to t+1}$; $\tilde{J}_i^t$ 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 $J_i^t$, $\widehat{J''}_i^t$ and $\widehat{J'}_i^t$, where i and t are positive integers.

**[0013]** It can be understood that a coordinate point set of pose key points of one human instance is used to represent the pose key points of the human instance. In addition, the $\widehat{J'}_i^t$ specifically takes the pose key points of human instances in the i-th initial human bounding box in the (t–1)-th image frame as a reference, and obtains the pose key points of the same human instance in the t-th image frame according to the motion status of the human instance in the (t–1)-th image frame reflected by $F^{t-1\to t}$ from the (t–1)th image frame to the t-th image frame, i.e., the position change in the i-th initial human bounding box in the (t–1)-th image frame from the (t–1)-th image frame to the t-th image frame. Similarly, the description of $\widehat{J''}_i^t$ is similar to that of $\widehat{J'}_i^t$, which will not be repeated herein. Furthermore, by synthesizing $\widehat{J'}_i^t$, $\widehat{J''}_i^t$ and $J_i^t$, i.e., by weighted average processing of corresponding coordinates of the three points, the initial pose key points of the human instances in the t-th image frame can be corrected to obtain a

more accurate detected pose key point.

**[0014]** In a possible implementation of the first aspect, before the recognizing poses of the human instances in the multi-human image based on the detected human bounding boxes of the multi-human image, the method further includes: generating identification information corresponding to the detected human bounding boxes of the human instances in each multi-human image frame, where the identification information is the following ID, and one piece of identification information is used for identifying the detected human bounding box of one human instance. It can be understood that the identification information corresponding to the detected human bounding box of the human instance specifically represents the human instance in the detected human bounding box. Specifically, in the process of moving each human instance in a plurality of multi-human image frames, the detected human bounding boxes where different human instances are located can be distinguished by the identification information, that is, different human instances can be distinguished.

**[0015]** In a possible implementation of the first aspect, the recognizing poses of the human instances in the multi-human image based on the detected human bounding boxes of the multi-human image includes: 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. It can be understood that in the process of detecting and obtaining detected pose key points of human instances in a plurality of multi-human image frames, different human instances can be distinguished based on the identification information of the detected human bounding boxes, thereby ensuring that the pose key points of the same human instance in adjacent images are corrected. This further improves the accuracy of pose key points (i.e., pose detection key points) of human instances and further improves the accuracy of pose estimation of human instances.

**[0016]** In a possible implementation of the first aspect, the method further includes: 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. It can be understood that, since the time domain information reflected by the position difference data between the plurality of multi-human image frames is used in the process of determining the pose key points of each human instance in each multi-human image frame, it can be ensured that the detected human bounding box and pose key points of each human instance in each multi-human image frame are completely detected, thus avoiding missing detection or incorrect detection. Therefore, the identification information can be used to accurately identify the detected human bounding box of each human instance in the plurality of multi-human image frames. That is, each human instance is identified, thereby implementing more accurate human tracking. For example, if the human instance "Tom" begins to appear in a certain multi-human image frame among the plurality of multi-human image frames, i.e., the identification information of the detected human bounding box where the human instance is located is "Tom", then as long as the human instance "Tom" appears in each multi-human image frame after this multi-human image, the identification information "Tom" of the detected human bounding box of this human instance "Tom" will be accurately generated in these multi-human images.

**[0017]** In a possible implementation of the first aspect, 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 backbone network of the first neural network model includes at least one of the following: ResNet152 network, ResNeXt101 network, or SeNet154 network. It can be understood that when a backbone network of the first neural network model includes the plurality of networks above, the first neural network model has all the functions of these networks. Specifically, the first neural network may be a human detection model below.

**[0018]** In a possible implementation of the first aspect, the initial pose key points of the human instances in each multi-human image frame are generated by a second neural network model, and a backbone network of the second neural network model includes at least one of the following: HRNet and SimpleNet. It can be understood that when a backbone network of the second neural network model includes the plurality of networks above, the second neural

network model has all the functions of these networks. Specifically, the second neural network may be a pose estimation model below.

**[0019]** In a second aspect, the embodiments of the present application disclose a video-based human pose estimation method, the method including: obtaining a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances; generating detected human bounding boxes of the human instances in each multi-human image frame; generating initial pose key points of the human instances in each multi-human image frame; 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, where the position difference data between the multi-human image and the previous image frame includes 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 includes 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.

**[0020]** In a possible implementation of the second aspect, 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 includes:

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:

$$\widehat{J'}_i^{\mathrm{t}}=J_i^{\mathrm{t}-1}+F^{\mathrm{t}-1\rightarrow\mathrm{t}},$$

$$\widehat{J''}_i^{t}=J_i^{t+1}-F^{t\to t+1},$$

$$\tilde{J}_i^{t}=\beta\cdot\widehat{J'}_i^{t}+\beta\cdot\widehat{J''}_i^{t}+(1-2\beta)\cdot J_i^{t},$$

where $J_i^{t-1}$ 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, $F^{t-1\to t}$ represents position difference data between the (t–1)-th image frame and the t-th image frame, $\widehat{J'}_i^{t}$ 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 $F^{t-1\to t}$; $J_i^{t+1}$ 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; $F^{t\to t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\widehat{J''}_i^{t}$ 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 $F^{t\to t+1}$; $\tilde{J}_i^{t}$ 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 $\beta$ represents a weight coefficient synthesized between $J_i^{t}$, $\widehat{J'}_i^{t}$ and $\widehat{J''}_i^{t}$, where i and t are positive integers.

**[0021]** In a possible implementation of the second aspect, the generating detected human bounding boxes of human instances in each multi-human image frame includes: 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, where the position difference data includes 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.

**[0022]** In a possible implementation of the second aspect, 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 includes: 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:

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

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

where $I_i^{t-1}$ 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, $F^{t-1\to t}$ 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 $F^{t-1\to t}$, $I_i^t$ 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 $\alpha$ represents a weight coefficient synthesized by $I_i^t$ and $\hat{I}_i^t$, where i and t are positive integers.

**[0023]** In a third aspect, the embodiments of the present application disclose a video-based human pose estimation apparatus, including: an image acquisition module, configured to obtain a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances; a human detection module, configured to generate initial human bounding boxes of the human instances in each multi-human image frame; a human correction module, configured to correct the initial human bounding boxes 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 detected human bounding boxes, where the position difference data includes a position difference of the human instances in the multi-human image between the previous image frame and the multi-human image; and a pose estimation module, configured to recognize poses of the human instances in the multi-human image based on the detected human bounding boxes of the multi-human image.

**[0024]** In a fourth aspect, the embodiments of the present application disclose a video-based human pose estimation apparatus, including: an image acquisition module, configured to obtain a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances; a human detection module, configured to generate detected human bounding boxes of the human instances in each multi-human image frame; 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; 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, where the position difference data between the multi-human image and the previous image frame includes 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 includes a position difference of the human instances in the multi-human image between the multi-human image and the next image frame; and 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.

**[0025]** In a fifth aspect, the embodiments of the present application disclose a readable medium, where 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 pose estimation method according to the first aspect or the second aspect.

**[0026]** In a sixth aspect, the embodiments of the present application disclose an electronic device, including: 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 pose estimation method according to the

first aspect or the second aspect.

## BRIEF DESCRIPTION OF DRAWINGS

**[0027]** FIG. 1 shows a schematic diagram of a target object according to some embodiments of the present application;

**[0028]** FIG. 2 shows a schematic diagram of an optical flow between two consecutive frames according to some embodiments of the present application;

**[0029]** FIG. 3 shows a diagram of a human pose estimation scene according to some embodiments of the present application;

**[0030]** FIG. 4 shows a schematic diagram of a composition architecture of a pose estimation scene according to some embodiments of the present application;

**[0031]** FIG. 5 shows a flowchart of a pose estimation method according to some embodiments of the present application;

**[0032]** FIG. 6 shows a schematic diagram of a bounding box output by a human detection model according to some embodiments of the present application;

**[0033]** FIG. 7 shows a schematic diagram of coordinate points of a human bounding box corresponding to a human instance according to some embodiments of the present application;

**[0034]** FIG. 8 shows a schematic diagram of a composition architecture of pose tracking according to some embodiments of the present application;

**[0035]** FIG. 9 shows a schematic diagram of distribution of human joint points and torsos output by a human pose estimation model according to some embodiments of the present application;

**[0036]** FIG. 10 shows a schematic diagram of human joint point coordinates detected from human instances of a multi-human image according to some embodiments of the present application;

**[0037]** FIG. 11 shows a schematic diagram of possible output categories and probabilities when a human body detection model detects a human instance according to some embodiments

of the present application;

**[0038]** FIG. 12 shows a video-based human pose estimation apparatus according to some embodiments of the present application;

**[0039]** FIG. 13 shows another video-based human pose estimation apparatus according to some embodiments of the present application;

**[0040]** FIG. 14 shows a block diagram of an electronic device according to some embodiments of the present application;

**[0041]** FIG. 15 shows a block diagram of a system on chip (SoC) according to some embodiments of the present application.

## DESCRIPTION OF EMBODIMENTS

**[0042]** Illustrative embodiments of the present application include, but are not limited to, a pose estimation method and apparatus, a medium, and an electronic device.

**[0043]** The video-based human pose estimation method in the present application is applicable to various video processing scenes, especially crowded scenes. For example, this scene is applied to robot vision processing, video surveillance, human action recognition, animal action recognition, etc. Specifically, for example, this scene can be used to determine whether the elderly on the road or living alone have falling poses, whether animals in zoos have abnormal or dangerous poses, and whether pedestrians in crowded scenes have pushing or trampling poses.

**[0044]** According to the pose estimation method provided in the embodiments of the present application, a plurality of image frames of a video to be processed can be obtained, bounding boxes of human bodies in the plurality of image frames can be detected by a human detection model for detecting human bodies, and then the plurality of image frames with human bodies detected can be detected again by a human pose estimation model for detecting human poses to obtain joint points of each human body in the plurality of image frames. Then, the method can be used to correct (i.e., smooth) the detected joint points of the human bodies

through the optical flow between the plurality of image frames, i.e., motion information of the human bodies based on the time domain in the plurality of image frames, thereby avoiding missing frames or dropping frames during processing of the plurality of image frames, and improving the accuracy of the finally obtained joint points of the human bodies, i.e., improving the accuracy of human pose estimation. In addition, the method can also be used to correct, after the human bounding boxes are detected from the plurality of image frames by the human detection model, the detected bounding boxes of human bodies through the optical flow between the plurality of image frames, so as to deal with missing frames or dropping frames in the plurality of image frames. This improves the accuracy of the finally detected human bounding boxes, and further improves the accuracy of subsequent human pose estimation.

**[0045]** For ease of understanding the technical solutions provided in the embodiments of the present application, the key terms used in the embodiments of the present application are first explained here:

**[0046]** Target object: with respect to a frame of image, the target object is an object that needs to be subjected to pose estimation. The setting of the target object may be different for different application scenes. FIG. 1 is a schematic diagram of a frame of image. If the target object is set as a human, there are two target objects in FIG. 1, i.e., a target object 1 and a target object 2. It can be understood that in a zoo scene, the target objects may be set as orangutans, tigers and other animals. In the embodiments of the present application, the target object being a human (i.e., human body) is taken as example. In particular, the human instances or human bodies herein refer to the human bodies in FIG. 1. To facilitate different descriptions, it is sometimes called "human instance" and sometimes called "human bodies".

**[0047]** Bounding box: a target detection box for indicating the position of the target object. The "target detection box" is also called the "bounding box". The target detection box can be a rectangular box, a circular box, or other shapes or an envelope line surrounding the target object. If the target detection box is a rectangular box, such as a rectangular box or a square box, the position information of the target detection box includes the coordinates of the point at the lower

left corner of the target detection box and the coordinates of the point at the upper right corner. If the target detection box is a circular box, the position information of the target detection box includes the center coordinates and radius length of the target detection box. In the embodiments of the present application, description is made by taking the target detection box being a rectangular box as an example. In particular, the target object being human is taken as example. The bounding box of the human body or the human bounding box herein refers to the position of the human body in the image.

**[0048]** Optical flow: it is used to define the movement of objects in an image sequence (such as images in a video stream), represented by a two-dimensional vector (or two-dimensional image), specifically used to represent the amount of movement of pixels representing the same object in a frame of the video image to the next frame. That is, the optical flow refers to the position difference of the target point, i.e., the change in the coordinates of the target point when the time interval between video frames is very small, i.e., between two consecutive frames of the video. In this way, the motion information of the object (such as the human body) between adjacent frames can be calculated by using the changes of the pixels in the image sequence in the time domain and the correlation between adjacent frames to find the correspondence between the previous frame and the current frame. For example, FIG. 2 is a schematic diagram of the optical flow between two consecutive frames (i.e., the (t–1)-th frame and the t-th frame) in a video stream, FIG. 2(a) shows the image of the (t–1)-th frame, FIG. 2(b) shows the image of the t-th frame, and FIG. 2(c) shows the optical flow from the (t–1)-th frame to the t-th frame. It can be seen from the figure that the optical flow is the displacement change in the time domain from the (t–1)-th frame to the t-th frame. In the embodiments of the present application, the optical flow is used to reflect the motion information of the human body between two consecutive frames in the video to be processed, i.e., the position difference (i.e., displacement) of the human body. The optical flow can be expressed by the formula:

$$F^{t-1\to t} = I^{t}(x+\Delta x, y+\Delta y) - I^{t-1}(x, y)$$

where $F^{t-1\to t}$ represents the optical flow from the (t–1)-th frame to the t-th frame,

$(x, y)$ represents a non-boundary coordinate value of the image pixel, $I^t$ represents the t-th image frame, and $I^{t-1}$ represents the (t–1)-th image frame. $\Delta x$, $\Delta y$ represents the position difference from the (t–1)-th frame to the t-th frame.

**[0049]** Joint points: in a human pose recognition scene, joint points represent preset key points of the human body, such as neck key points, elbow key points, wrist key points, shoulder key points and head key points. In the embodiments of the present application, the joint points recognized in the image are represented by corresponding coordinate points of the joint in the image.

**[0050]** For ease of description of the technical solution of the present application, the following description takes human pose estimation under video surveillance as an example.

**[0051]** FIG. 3 shows a human pose estimation scene under video surveillance of a shopping mall according to the embodiments of the present application. As shown in FIG. 3, this scene includes three cameras 301 and a video processing device 300. It can be understood that a pose estimation scene applicable to the technical solutions of the present application may include any number of cameras, not limited to three.

**[0052]** The camera 301 is configured to capture or shoot a video stream containing human bodies, and the video processing device 300 is configured to estimate poses of the human bodies in the obtained video stream containing the human bodies, such as determining that the poses of the human bodies are lying, sitting, standing, walking or walking up stairs, etc.

**[0053]** In some embodiments, the cameras 301 directly transmit the captured video stream to the video processing device 300. In other embodiments, the cameras 301 may also transmit the captured video stream to a designated storage device for storage, and then the video processing device 300 obtains the video stream from the storage device, which is not limited in the embodiments of the present application.

**[0054]** In some embodiments of the present application, the video processing device 300 is a computer device with a specific image or video processing capability, such as a personal computer (PC), a notebook computer, or a server. The server can be an independent physical

server, or a server cluster consisting of a plurality of physical servers, or a server that provides basic cloud computing services such as cloud database, cloud storage, and CDN. The scale of the server can be planned according to the number of video streams to be processed, which is not limited in the embodiments of the present application.

**[0055]** As an example, in the scene of performing pose estimation on the human body in video surveillance, the cameras 301 may be surveillance cameras disposed in places such as shopping malls, roads, or subway entrances/exits for shooting video streams of pedestrians in these places. In practical applications, the foregoing pose estimation scene may include a larger number of cameras 301, for example, cameras 301 disposed at various positions of each floor in a shopping mall.

**[0056]** In some embodiments, the camera 301 and the video processing device 300 are independent devices. For example, the camera is a surveillance camera and the video processing device 300 is a server that obtains video streams from the surveillance camera and processes the video streams. In other embodiments, the camera 301 and the video processing device 300 are different components in the same device, such as a camera and a processor of a mobile phone, respectively.

**[0057]** In addition, the camera 301 and the video processing device 300 may be communicatively connected through one or more networks. The network can be a wired network or a wireless network. For example, the wireless network can be a mobile cellular network, or a wireless-fidelity (Wi-Fi) network, and certainly, it can also be other possible networks, which is not limited in the embodiments of the present application.

**[0058]** In some embodiments of the present application, the video-based pose estimation method can be performed by a pose estimation composition architecture, for example, the composition architecture runs in the foregoing video processing device 300.

**[0059]** FIG. 4 shows a schematic diagram of a composition architecture of a pose estimation scene according to an embodiment of the present application. As shown in FIG. 4, the composition architecture 40 includes a human detection model 41, an optical flow smoothing

branch 42, a multi-person tracking branch 43, a pose estimation model 44, an optical flow smoothing branch 45, a plurality of consecutive image frames 46, a plurality of consecutive image frames 47 of human joint points and an optical flow 48 between the plurality of consecutive image frames.

**[0060]** Each part of the composition architecture 40 of the pose estimation scene is specifically described below.

**[0061]** The human detection model 41 is configured to detect human bodies in the plurality of input image frames (i.e., an image sequence). Specifically, the human detection model 41 first classifies objects in each image frame to obtain a region where the human instances are located, and then determines the position of the region where the human instances are located in each image frame, and marks the human bounding boxes at the corresponding positions. The related content of the training of the human detection model 41 will be described in detail below, and will not be repeated here.

**[0062]** The optical flow smoothing branch 42 is configured to correct the bounding boxes of the human bodies generated by the human detection model 41, detect a plurality of image frames of the bounding boxes of the human bodies for the human detection model 41, and optimize the bounding boxes of the human bodies by using an optical flow 48 between the plurality of image frames, i.e., adjust the positions of the bounding boxes of the human bodies, i.e., adjust the positions of the detected human bodies in the plurality of image frames.

**[0063]** The multi-person tracking branch 43 is configured to distinguish human instances in the images and implement multi-person tracking of different image frames. Specifically, for a plurality of image frames with detected human bounding boxes, such as a plurality of image frames which pass through the human detection model 41 and the optical flow smoothing branch 42 in sequence and are output, an ID is added for each human instance in the images based on the bounding boxes of the human bodies generated by the plurality of output image frames, deep appearance features of the human instances in the bounding boxes and the position information of the human instances in the images.

**[0064]** The pose estimation model 44 is configured to estimate poses of the human bodies in the plurality of input image frames. Specifically, with respect to a plurality of image frames that sequentially pass through the human detection model 41, the optical flow smoothing branch 42 and the multi-person tracking branch 43, the pose estimation model 44 outputs heat map features of human pose features, then obtains human joint point coordinates and torso based on the heat map of human pose features, and estimates human poses through the joint points and torso, for example, estimates whether the pose of a human instance is walking up stairs, sitting or bending, etc. The related content of the training of the pose estimation model will be described in detail below, and will not be repeated here.

**[0065]** The optical flow smoothing branch 45 is configured to correct human joint points output by the pose estimation model, detect human joint point coordinates for the pose estimation model 44, and optimize human joint point coordinates by using the optical flow 48 between the plurality of image frames, i.e., adjust the positions of these joint points, i.e., adjust the positions of the detected human bodies in the plurality of image frames.

**[0066]** A specific solution in the present application is described below with reference to FIG. 1 to FIG. 5. FIG. 5 shows a flowchart of a pose estimation method according to an embodiment of the present application. Specifically, this solution includes the following steps.

**[0067]** Step 502: The video processing device 300 obtains a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances.

**[0068]** In some embodiments of the present application, the video to be processed in step 502 may be a surveillance video shot by the camera 301, or the video to be processed may be a video uploaded by the user to a video platform or short video social software. The video to be processed may also be a video in the test set, that is, the pose estimation method of the present application is used to perform human pose estimation on the video in the test set. The video-based pose estimation method of the present application is described herein by mainly taking the human pose estimation of a video in a crowded scene as an example.

**[0069]** In other embodiments of the present application, step 502 may first obtain the video

to be processed, and then preprocess the video to be processed in real time, such as performing a framing operation to obtain a plurality of image frames of the video to be processed. For example, video framing software or video screenshot software can be used to perform a framing operation on the video to be processed. For example, the video processing device 300 performs the framing operation on the video to be processed through the videocapture function of an Opencv library of the software, to generate a plurality of consecutive image frames 46.

**[0070]** For example, the duration of the video to be processed is 10 minutes, and during the framing operation according to the preset frame rate, assuming that the frame rate is 60 fps, the number of images obtained after the framing and parsing of the video to be processed should be 60*10*60=36,000 images, and it can be understood that 36,000 images are obtained after 10 minutes of video framing. "*" indicates multiplication.

**[0071]** It can be understood that the plurality of consecutive image frames 46 of the video to be processed are used to be input into the human detection model 41 for human detection.

**[0072]** Step 504: The human detection model 41 generates initial human bounding boxes of human instances in each multi-human image frame.

**[0073]** The human detection model 41 can achieve three tasks. The first task is a classification task, which is to traverse each input image frame and classify the human bodies and non-human bodies of each image frame. The second task is a regression task, which is to determine the position of the human region in each image frame, and annotate the target bounding box at the corresponding position. The third task is a coordinate point positioning task, which is to determine the position of the coordinate point of a bounding box of the human body, where the bounding box can be a rectangular box surrounding the region where the human body is located. For example, for the image shown in FIG. 6, the human detection model 41 can generate a bounding box 1 of the human body for the human instance 1, and a bounding box 2 of the human body for the human instance 2. It can be understood that the three tasks above are complementary to each other. For example, determining the position of the bounding box coordinates of the human body is used to determine the position of the human body in the

image.

**[0074]** Specifically, the bounding box of a human body can be represented by a coordinate point set consisting of the coordinate points of a plurality of vertices thereof, for example, it is represented by the coordinates at the lower left corner and the upper right corner of the bounding box of the human body. It can be understood that, in the embodiments of the present application, the human detection model 41 detects a human instance and determines the bounding box of the human body in the region where the human body is located, that is, determines the coordinate point set of the bounding box of the human body. That is, step 504 specifically detects and obtains the coordinate point set of the bounding box of each human body in the plurality of image frames.

**[0075]** In the embodiments of the present application, the process of human detection performed by the human detection model 41 on multi-human images in the plurality of input consecutive image frames 46 includes: obtaining the position of each human body in the image, and predicting the bounding box of each human body for the region where the human body is located. Specifically, each image frame can be input into the pre-trained human detection model 41 for human detection, the region of the human body in the image is obtained, and the human detection model 41 predicts the bounding box of the human body for the region where each human body is located when human detection is performed on each image frame. The human detection model 41 can be a pre-trained deep learning convolutional neural network. The human detection model 41 is trained using a large-scale training sample set consisting of images with human instance annotations. During model training, it is necessary to enable the model to accurately predict the region where the human body is located in each image frame, so that the human detection model 41 outputs the predicted bounding box of the human body more accurately.

**[0076]** In the embodiments of the present application, taking the video to be processed being a video in a crowded scene as an example, for each of the consecutive image frames 46 after framing of the video to be processed, multi-human images are sequentially input into the

trained human detection model 41 to obtain initial human bounding boxes output by the human detection model 41.

**[0077]** Step 506: The optical flow smoothing branch 42 corrects 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, where the position difference data includes a position difference of the human instance in the multi-human image between the previous image frame and the multi-human image.

**[0078]** In the embodiments of the present application, the optical flow smoothing is performed on the bounding boxes of all human bodies in the multi-human image. The optical flow smoothing is to correct an initial human bounding box of the multi-human image based on position difference (i.e., displacement) data between the multi-human image and the previous image frame of the multi-human image, to obtain the detected human bounding box, where the position difference data includes a position difference of the human instance in the multi-human image between the previous image frame and the multi-human image. It can be understood that predicting the bounding box of the human body in the current frame based on the position difference in the time domain from the previous frame to the current frame can effectively avoid missing or dropping frames.

**[0079]** Specifically, for the same human body, the detected human bounding box of the region where the human body is located is obtained according to the initial human bounding box of the current frame output by the human detection model 41 and the initial human bounding box of the previous frame output by the human detection model 41, combined with the optical flow from the previous frame to the current frame. According to the optical flow smoothing processing, that is, according to the motion information of the human body in the plurality of image frames based on the time domain, the detected initial bounding box is corrected (i.e., smoothed), to avoid missing or dropping frames when processing the plurality of image frames, thereby improving the accuracy of the finally detected human bounding box,

that is, improving the accuracy of human pose estimation.

**[0080]** It can be understood that, in the embodiments of the present application, the optical flow smoothing is performed on the bounding box of the human body detected by the human detection model 41 in the current frame in step 504, specifically, smoothing is performed on the coordinate points in the coordinate point set of the bounding boxes of these human bodies to obtain the detection coordinate point set of the bounding box of each human body, that is, the detected human bounding box of each human body in the current frame is obtained. The detected human bounding box of a human body in the embodiments of the present application is a human bounding box optimized through the optical flow smoothing. Specifically, the coordinate points in the detection coordinate point set of the bounding box of a human body are coordinate points optimized through the optical flow smoothing. The detection coordinate point set of the detected human bounding box of a human body obtained by the optical flow smoothing branch 42 is used for subsequent input to the multi-person tracking branch 43 and the pose estimation model 45 for processing.

**[0081]** Specifically, taking the optical flow smoothing of the bounding box of the i-th human body (annotated as the human bounding box I1) in the t-th image frame of the processed video through the following formula as an example, it is illustrated that the optical flow smoothing branch 42 performs optical flow smoothing on the bounding box of the human body detected by the human detection model 41 in the embodiments of the present application. For example, the t-th image frame is the multi-human image in step 506.

**[0082]** FIG. 7 is a schematic diagram of a coordinate point of the human bounding box I1 corresponding to the human instance I1 in the t-th multi-human image frame. Specifically, as shown in the figure, the coordinate point set $\tilde{I}_i^k$ detected by using the human bounding box I1 includes a point n $(x_1, y_1)$ at the upper right corner of the human bounding box and a point m $(x_2, y_2)$ at the lower left corner of the human bounding box.It can be understood that by obtaining the point n $(x_1, y_1)$ at the upper right corner and the point m $(x_2, y_2)$ at the lower left corner of the human bounding box I1, the region of the human instance I1 in the multi-

human image can be obtained.

**[0083]** In a possible implementation, the embodiments of the present application calculate the detected coordinate point set $\tilde{I}_i^k$ obtained through optical flow smoothing of the human bounding box I1 by using the following formula (1):

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

**[0084]** In the formula, in $\tilde{I}_i^t$, $\hat{I}_i^t$ and $I_i^t$, $t$ represents the $t$-th image frame, and i represents the bounding box of the i-th human body in the t-th image frame.

**[0085]** $\alpha$ represents a weight coefficient, which is an empirical parameter; in most cases, the value is 0.5.

**[0086]** $\hat{I}_i^t$ represents the coordinate point set of the human bounding box I1 in the t-th image frame obtained through prediction according to the optical flow from the (t–1)-th frame to the t-th frame.

**[0087]** $I_i^t$ represents the coordinate point set of the initial human bounding box I1 in the $t$-th image frame output by the human detection model 41.

**[0088]** It can be understood that the formula (1) can be used to determine the detected coordinate point set $\tilde{I}_i^t$ of the detected human bounding box I1 of the t-th image frame according to the coordinate point set $\hat{I}_i^t$ of the human bounding box I1 of the t-th image frame predicted by optical flow from the (t–1)-th image frame to the t-th image frame and the coordinate point set $I_i^t$ of the initial human bounding box I1 of the $t$-th frame output by the human detection model 41.

**[0089]** In addition, some embodiments of the present application can implement "the coordinate point set $\hat{I}_i^t$ of the human bounding box I1 of the t-th image frame predicted according to the optical flow from the (t–1)-th frame to the t-th frame" by using Formula (2):

$$\hat{I}_i^t = I_i^{t-1} + F^{t-1\to t} \quad (2)$$

**[0090]** In the formula, $F^{k-1\to k}$: the optical flow of the joint coordinate points of the human bounding box I1 from the (t–1)-th frame to the t-th frame.

**[0091]** $I_i^{t-1}$ represents the coordinate point set of the initial human bounding box I1 in the

(t–1)-th frame output by the human detection model 41.

**[0092]** Step 508: The multi-person tracking branch 43 generates identification information corresponding to the detected human bounding boxes of human instances in each multi-human image frame, where one piece of identification information is used for identifying the detected human bounding box of one human instance.

**[0093]** In the prior art, a degree of overlap of the bounding boxes or a pose similarity between the previous frame and the current frame is applied to determine whether the bounding boxes come from one human instance, so as to track poses of the same human instance in adjacent frames in a plurality of multi-human image frames. However, due to the serious occlusion and overlap of human bodies in crowded scenes, it is impossible to track poses of the same human instance in adjacent frames by the prior art.

**[0094]** The multi-person tracking branch 43 generates identification information corresponding to the detected human bounding boxes of the human instances in each multi-human image frame, where the identification information is the following ID, and one piece of identification information is used for identifying the detected human bounding box of one human instance. It can be understood that the bounding box of the human instance is associated with the human instance. in the process of moving each human instance in a plurality of multi-human image frames, the detected human bounding boxes where different human instances are located can be distinguished by the identification information, that is, different human instances can be distinguished.

**[0095]** In the embodiments of the present application, compared with the prior art, the present application adds an ReID model for recognizing the deep appearance features of human instances. The ReID model is an Coarse-to-Fine Pyramidal model.

**[0096]** With reference to FIG. 4 and FIG. 8, the multi-person tracking branch 43 includes the ReID model and a DeepSORT algorithm, where the DeepSORT algorithm is used to assign a unique ID to the bounding box of the human body with reference to the position of the bounding box of the human body and the appearance features of the human body in the multi-

person image. Specifically, the pre-trained ReID model is used to extract the deep appearance features of the human body in each image frame, and the appearance features extracted by the ReID model and the position information of the bounding box of the human body are input into the DeepSORT algorithm, and each human instance in the image is assigned a unique ID. For example, as shown in FIG. 8, a multi-human image 82 generates bounding boxes for marking human bodies for Human Detection 81. The human instance in the bounding box of the human body of the multi-human image 82 is subjected to feature extraction by the ReID model, and the extracted features n×256 are output to the DeepSORT algorithm, where n represents the number of human instances contained in the multi-human image 82 and 256 represents the long dimension of the image feature vector of the human instance. Among them, the value of the long dimension is variable according to the dimension requirements for extracting the depth appearance feature.The multi-human image 82 outputs the position information n×4 of the bounding box of the human body to the DeepSORT algorithm, where n represents the number of human instances contained in the multi-human image 82 and 4 represents the horizontal and vertical coordinate points of the lower left corner and upper right corner of the bounding box of the human body. It can be understood that, on the one hand, for each multi-human image frame, the ReID model extracts the deep appearance features of the human body in the multi-human images, and the same human body instance is assigned a unique ID (i.e., identification information) in the multi-human images of different frames. It can be understood that assigning an ID to a human instance means assigning an ID to the bounding box of the human instance. On the other hand, when a new human instance appears in the multi-human image, the ReID model can distinguish the new human instance from the human instance assigned with an ID, and the multi-person tracking branch 43 assigns a unique ID to the new human instance. Compared with the prior art, the ReID model pays more attention to people's appearance, so it is widely used in multi-person tracking of various scenes.

**[0097]** In addition, for example, the "Human Detection" 81 may consist of a human detection model 41 and an optical flow smoothing branch 42 shown in FIG. 4. Or, in FIG. 8,

the optical flow smoothing branch 42 is omitted for ease of description, but the bounding box of the human body in the image output via the "Human Detection" 81 is the bounding box of the human body after optical flow smoothing.

**[0098]** In the embodiments of the present application, in addition to adding the ReID model to recognize the deep appearance features of human instances and distinguish human instances, the following two aspects are also modified.

**[0099]** (1) Optimal parameter search

**[00100]** In order to obtain the optimal parameter set to make the tracking method run normally on the test set, in the present application, grid search is performed on the training set. Search space is designed and optimized for five parameters, i.e., the maximum cosine distance (max-cos-dis), the number of object appearance history to keep (nn-budget), the max age of a track (max-age), the number of frames for a tentative track to be confirmed (nn-init) and the maximum IoU distance (max-iou-dis), respectively. The optimal parameter design is selected through multi-target tracking evaluation indicators. The optimal parameters on the training set and the test set and the human tracking performance results of the multi-person tracking branch 43 under the optimal parameters are shown in Table 1. The performance in the table is also the performance of the final multi-person tracking branch 43 used in pose tracking in the embodiments of the present application. The multi-target tracking evaluation indicators are respectively as follows: Multiple object tracking accuracy (MOTA): When its value is larger, the performance is better. Multiple object tracking precision (MOTP): When its value is larger, the performance is better. The total number of false positives (FP): When its value is smaller, the performance is better. The total number of false negatives (FN): When its value is smaller, the performance is better. The total number of identity switches (ID SW): When its value is smaller, the performance is better.

| Table 1 | | | | |
|---|---|---|---|---|
| Parameter | | | | |
| max-cos-dis | nn-budget | max-age | nn-init | max-iou-dis |
| 0.3 | 256 | 30 | 3 | 0.7 |
| Performance on the training set | | | | |
| MOTA | MOTP | FP | FN | ID Sw |
| 87.7 | 0.131 | 24839 | 61677 | 10175 |
| Performance on the test set | | | | |
| MOTA | MOTP | FP | FN | ID Sw |
| 61.85 | 0.78 | 3710 | 21211 | 1657 |

**[0100]** (2) Combination of functions of the ReID model

**[0101]** In the embodiments of the present application, the ReID model can extract the deep appearance features of the human body in an image. On the one hand, in an image containing a plurality of human instances, different human instances can be distinguished based on the deep appearance features of the human body. On the other hand, based on different image frames, human pose tracking can be implemented by determining that the same human appearance features are of the same human instance.

**[0102]** Specifically, the DeepSORT algorithm is combined with different strategies of the functions of the ReID model. Strategy A: linear combination of original features and features extracted from horizontally flipped humans; strategy B: linear combination of ReID features extracted by using models trained with different scales. Table 2 lists the test results of multi-person tracking with different combination strategies of the DeepSORT algorithm and the ReID model, where the bounding box of the tracked human body is generated by a baseline model, and the baseline model is the human detection model 41 with a Faster RCNN framework of ResNet50. Horizontal flip represents the strategy A, and Scale represents the strategy B. "+" represents the combination of the ReID model or the strategy A or the strategy B with the

DeepSORT algorithm.

**[0103]** It can be seen from Table 2 that through combination of a plurality of multi-target tracking evaluation indicators, the performance test results are the best when the DeepSORT algorithm is combined with the ReID model, the strategy A and the strategy B. Specifically, MOTA is increased from 27.11 to 53.96. FP is increased from 5894 to 5922, FN is decreased from 42668 to 24501, and ID Sw is decreased from 2220 to 1657.

Table 2

| Combination strategy | MOTA | FP | FN | ID Sw |
|---|---|---|---|---|
| DeepSORT | 27.11 | 5894 | 42668 | 2220 |
| DeepSORT +ReID model | 53.46 | 5622 | 24836 | 1963 |
| DeepSORT +ReID model +Horizontal flip | 53.5 | 5653 | 24798 | 1949 |
| DeepSORT +ReID model +Horizontal flip +Scale | 53.96 | 5922 | 24501 | 1657 |

**[0104]** Step 510: The human pose estimation model 44 recognizes 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.

**[0105]** In the embodiments of the present application, the human pose estimation model 44 is designed in the top-down mode, that is, the human detection model 41 obtains the position of each human instance in the multi-human image and outputs the bounding box of the human body, and then estimates joint points of the human instances in the bounding box of the human body one by one. Specifically, the human pose estimation model 44 traverses each input image frame, and an intermediate layer of the model outputs heat map features of pose features. Then, the human joint points and torso can be obtained based on the heat map of pose features, and the human poses can be estimated through the human joint points and torso. The torso represents the connecting line of adjacent joint points of the human body.

**[0106]** The human pose estimation model 44 is a pre-trained deep learning convolutional

neural network. In the embodiments of the present application, two most advanced single-person pose estimation models HRNet and SimpleNet are used as basic networks for generating pose prediction. Different from general high-to-low and low-to-high resolution modes, HRNet can keep high-resolution representation in the whole process and fuse multi-resolution representation. SimpleNet is a simple and effective model, which only consists of the backbone ResNet reducing the resolution of feature map and several deconvolution layers constituting post prediction. In addition, in the present application, an FPN structure is inserted into SimpleNet to enhance the performance of small instances. Finally, the present application fuses the results of the two models by averaging the heat maps generated by the two models. For example, as shown in FIG. 8, the pose estimation model 44 includes two basic networks: SimpleNet and HRNet. Specifically, through two networks, SimpleNet and HRNet in the pose estimation model 44, pose t shown in FIG. 8 can be output. Pose t represents initial joint points of all human bodies in the t-th image frame.

**[0107]** Specifically, FIG. 9 is a schematic diagram of distribution of human joint points and torsos output by a human pose estimation model 44. Pose features heat maps obtained by the human pose estimation model include joint point heat maps and torso heat maps. In FIG. 9, the left diagram is a schematic diagram of human joint point distribution, there are 14 joint points, and the 14 joint points in the left diagram of FIG. 9 correspond to 14 joint point heat maps. The right diagram in FIG. 9 is a schematic diagram of torso distribution, there are 13 torsos, each torso is a connecting line of adjacent joint points, and 13 torsos in the right diagram in FIG. 9 correspond to 13 torso heat maps. It can be understood that the number of joint points used for human pose estimation may be 14, 17 or 18. As long as the human poses can be estimated, no restrictions are imposed on the number of human joint points and the positions of joint coordinate points.

**[0108]** Specifically, the poses of one human instance are represented by a coordinate point set consisting of a plurality of coordinate points, for example, by coordinates of 14 joint points of the human instance. It can be understood that in the embodiments of the present application,

the human pose estimation model 44 estimates the human pose by determining the relative positions of the joint points of the human instance, i.e., by determining a coordinate point set of the human instance. That is, in step 510, the human pose estimation of a plurality of image frames is implemented by determining the joint points of the human instance.

**[0109]** In the embodiments of the present application, a plurality of image frames with the tag ID generated by the multi-person tracking branch 43 are input into the human pose estimation model 44. The process of outputting human joint points and torso by the human pose estimation model 44 includes: obtaining a joint coordinate point set of each human body, and predicting poses for each human body based on the joint coordinate point set of the human body. Specifically, for each image frame, when pose estimation is performed on each human instance with tag ID, each image frame can be input into a pre-trained pose estimation model for pose estimation, and the pose estimation model 44 can output a joint coordinate point set of the initial human body and predicted poses. It can be understood that the pose estimation model is obtained through training based on a large-scale pose training sample set marked with human joint points, so that the model can predict the human poses in the bounding box of the human body through the joint coordinate point set of the human body.

**[0110]** Step 512: The optical flow smoothing branch 45 corrects 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.

**[0111]** In the embodiments of the present application, in order to further overcome the missing frame or dropped frame in the pose estimation process and reduce the tracking missing rate of the human pose estimation of the video to be processed, the optical flow smoothing branch 45 corrects the initial pose key points of 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, and obtains the poses of the detected human instances, where the position difference data between the multi-human image and the next image frame of the multi-human image includes a position difference of the human instances in the multi-human image between the multi-human image and the next image frame. Obtaining the poses of the detected human instance can be understood as smoothing of the joint coordinate points of the same human body in different frames, and obtaining of the poses of the detected human instances in the region where the human body is located, i.e., obtaining of a detected coordinate point set of human joint points.

**[0112]** It can be understood that, in the embodiments of the present application, if the human instances have the same ID in the previous frame and the next frame, and their confidence level scores are higher than a predetermined threshold, optical flow smoothing is performed on the initial joint points of the human instance output by the human pose estimation model 44 in the current frame in step 510. Specifically, smoothing is performed on the coordinate points in the coordinate point set of these joint points to obtain the detected joint coordinate point set of each human instance. That is, the positions of detected joint points of each human instance in the current frame are obtained. The relative positions of the detected joint points of one human instance in the embodiments of the present application are a joint coordinate point set optimized through optical flow smoothing. The detected joint coordinate point set of a human instance is processed by the optical flow smoothing branch 45, and the detected joint coordinate point set according to the human instances is used for the final human pose estimation.

**[0113]** Specifically, it is taken as an example that the joint coordinate point set corresponding to the human instance whose identification information is Idi in the t-th multi-human image frame in the processed video is subjected to the optical flow smoothing by using the following formula. In the embodiments of the present application, the optical flow smoothing branch 45 performs optical flow smoothing on the human joint coordinate point set obtained by the human pose estimation model 44. For example, as shown in FIG. 8, the t-th

image frame outputs image pose t through the human pose estimation model 44, the (t–1)-th image frame outputs image pose t-1 through the human pose estimation model 44, and the (t+1)-th image frame outputs image pose t+1 through the human post estimation model 44. It can be understood that the initial joint coordinate point set of the human instance in the image pose t-1 is combined with the optical flow from the (t–1)-th frame to the t-th frame to predict the joint coordinate point set of the human instance in the t-th frame. The initial joint coordinate point set of the human instance in the image pose t+1 is combined with the optical flow from the t-th frame to the (t+1)-th frame to predict the joint coordinate point set of the human instance in the t-th frame. The detected joint coordinate point set Final pose t is generated by weighting the joint coordinate point set predicted by the image pose t-1, the joint coordinate point set predicted by the image pose t+1 and the joint coordinate point set of the image pose t. It can be understood that the detected joint coordinate point set is used for human pose estimation in the embodiments of the present application.

**[0114]** Specifically, for example, FIG. 10 is a schematic diagram of detected human joint point coordinates of a human instance IDi in the t-th image frame. Specifically, as shown in FIG. 10, taking the detected human joint point coordinate set obtained by human instance IDi as an example, for example, the detected human joint coordinate point set of the human instance IDi in the t-th image frame is $\tilde{J}_i^t$, $\tilde{J}_i^t$ is a joint coordinate point set containing 14 coordinate points, namely head A $(x_1', y_1')$, neck B $(x_2', y_2')$, left shoulder C $(x_3', y_3')$, right shoulder D $(x_4', y_4')$, left elbow E $(x_5', y_5')$, right elbow F $(x_6', y_6')$, left hand G $(x_7', y_7')$, right hand H $(x_8', y_8')$, left waist K $(x_9', y_9')$, right waist L $(x_{10}', y_{10}')$, left knee O $(x_{11}', y_{11}')$, right knee P $(x_{12}', y_{12}')$, left foot Q $(x_{13}', y_{13}')$, and right foot R $(x_{14}', y_{14}')$. The detected joint coordinate point set $\tilde{J}_i^k$ is calculated by Formula 3:

$$\tilde{J}_i^t = \beta \cdot \hat{J}'^t_i + \beta \cdot \hat{J}''^t_i + (1 - 2\beta) \cdot J_i^t \quad (3)$$

where t in $\tilde{J}_i^t$, $J_i^t$, $\hat{J}''^t_i$ and $\hat{J}'^t_i$ represents the t-th frame, and i represents the human instance Idi.

**[0115]** β represents the weight coefficient synthesized between $J_i^t$, $\hat{J}''^t_i$ and $\hat{J}'^t_i$, and the

coefficient is an empirical parameter.

**[0116]** $\widehat{J'}_i^t$ represents a human joint coordinate point set of a human instance ID3 in the t-th image frame predicted by the optical flow from the (t–1)-th frame to the t-th frame.

**[0117]** $\widehat{J''}_i^t$ represents a human joint coordinate point set of a human instance ID3 in the t-th image frame predicted by the optical flow from the t-th frame to the (t+1)-th frame.

**[0118]** $J_i^t$ represents the initial human joint coordinate point set of the human instance ID3 in the t-th image frame output by the pose estimation model.

**[0119]** The human joint coordinate point set $\widehat{J'}_i^t$ of the human instance ID3 in the t-th image frame predicted by the optical flow from the (t–1)-th frame to the t-th frame is calculated by using Formula 4:

$$\widehat{J'}_i^t = J_i^{t-1} + F^{t-1\to t} \tag{4}$$

where $t-1$ in $J_i^{t-1}$ represents the $(t-1)$-th image frame, and i represents the i-th human body.

**[0120]** $J_i^{t-1}$ represents the initial human joint coordinate point set of the human instance ID3 in the (t–1)-th image frame output by the pose estimation model.

**[0121]** $F^{t-1\to t}$ represents optical flow of human joint coordinate points of the human instance ID3 from the (t–1)-th frame to the t-th frame.

**[0122]** The human joint coordinate point set $\widehat{J''}_i^t$ of the human instance ID3 in the t-th image frame predicted by the optical flow from the t-th frame to the (t+1)-th frame is calculated by using Formula 5:

$$\widehat{J''}_i^t = J_i^{t+1} - F^{t\to t+1} \tag{5}$$

where $t+1$ in $J_i^{t+1}$ represents the $(t+1)$-th image frame, and i represents the i-th human body.

**[0123]** $J_i^{t+1}$ represents the initial human joint coordinate point set of the human instance ID3 in the (t+1)-th image frame output by the pose estimation model.

**[0124]** $F^{t\to t+1}$ represents optical flow of human pose coordinate points of the human instance ID3 from the t-th frame to the (t+1)-th frame.

**[0125]** It can be understood that the pose of a human instance in a multi-human image is recognized based on the detected human joint points in the multi-human image.

**[0126]** In addition, it should be noted that in the embodiments of the present application, the human instances are tracked 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. It can be understood that each human instance is assigned a unique ID by extracting the depth appearance features of the human instance and combining the position information of the human instance. On the one hand, the same human instance can be determined by confirming the same appearance features and position information of instances in different frames; and on the other hand, different human instances can be distinguished in one image frame by confirming different deep appearance features and position information of human instances. Therefore, optical flow smoothing is performed on a plurality of image frames to determine the more accurate human pose, so that multi-person tracking can be implemented in the plurality of image frames, especially the tracking of newcomers and the tracking of people who leave from the image halfway can be implemented.

**[0127]** In the posture recognition solution of the present application, the human body in the multi-human image is detected by the human detection model 41 to generate a bounding box of the human body. Specifically, the function and training process of the model are as follows:

**[0128]** (1) Training purpose of the human detection model 41

**[0129]** In the embodiments of the present application, test experiment is performed on a video to be detected in a crowded scene by using the human detection model 41, and the performance of the human detection model 41 is tested through two indicators, i.e., average precision (AP) and MMR. The AP reflects the accuracy and recall rate of the detection results. The value of the MMR is between 0.01 and 100. The MMR refers to the logarithmic average

loss rate of false positive per image (FPPI). The MMR is very sensitive to false positives (FP), especially FP with high confidence level may seriously damage the MMR ratio. In the embodiments of the present application, the larger the AP and the smaller the MMR is, the better the performance of the human detection model 41 is.

**[0130]** (2) Selection of a model framework of the human detection model 41

**[0131]** In the prior art, for the detection of human instances in the video to be detected in a crowded scene, a detection framework in the human detection model 41 may be a one-stage detection framework, for example, the one-stage framework may be an SSD framework, a YOLO framework, etc. It can also be a two-stage framework, where the one-stage detection framework detects human recognition and human positioning together, which can run in real time and has a high recognition speed, but has lower accuracy than the two-stage framework. The two-stage detection framework completes the human recognition and the human positioning separately in two steps, and the recognition accuracy is high.

**[0132]** Therefore, in the embodiments of the present application, for the detection of human instances of the video to be detected in a crowded scene, a two-stage framework is adopted, such as cascade-region convolutional neural networks (cascade RCNN), faster-region convolutional neural networks (Faster RCNN), etc. The backbone network of the detection framework can be a ResNet152 network, a ResNeXt101 network, or a SeNet154 network.

**[0133]** The framework selection results in the human detection model 41 are evaluated by using table 3 below.

**[0134]** Table 3 shows the experimental results of the human detection model 41 using different backbone networks and detection frameworks as well as feature-pyramid networks (FPN) to perform human detection on the video to be detected in HIE validation set. The human detection models 41 of any components in the table are obtained through training based on the training sample set as the HIE data set. The "+" in the table represents the constituent parts (or components) of the human detection model 41. For example, ResNet152+Faster RCNN+FPN means that the detection model consists of the backbone network ResNet50, the

framework Faster RCNN and FPN. It should be noted that a baseline model is a human detection model 41 with the Faster RCNN framework of ResNet50. Other detection models in the table search the baseline model for hyper-parameters, and then apply same to a larger backbone or framework.

**[0135]** It can be seen from Table 3 that the performance of the detection model consisting of the backbone network ResNet152 or ResNeXt101 and the framework Cascade RCNN and FPN is significantly improved compared to the baseline model. For example, the detection model consisting of the backbone network ResNet152, the framework Cascade RCNN and FPN has the AP increased by 9.38 and the MMR reduced by 11.46 compared to the baseline model. The detection model consisting of the backbone network ResNeXt101, the framework Cascade RCNN and FPN has the AP increased by 9.38 and the MMR reduced by 11.46 compared to the baseline model. First, with the addition of FPN, the model performance has the AP increased by 5.64 and the MMR reduced by 5.84. Secondly, in the selection of the model backbone network: ResNet152 has a significant improvement in model performance compared to ResNet50 or SENet154. ResNeXt101 has a significant improvement in performance compared to SENet154. Thirdly, in the selection of the model framework network, the performance of Cascade RCNN has the AP increased by 1.79 and the MMR reduced by 2.33 compared to Faster RCNN.

Table 3

| Method or model | AP(%) | MMR(%) |
|---|---|---|
| Baseline（ResNet50+Faster RCNN） | 61.68 | 74.01 |
| ResNet152+Faster RCNN | 67.32 | 68.17 |
| ResNet152+Faster RCNN+FPN | 69.77 | 64.83 |
| SENet154+Faster RCNN+FPN | 65.77 | 68.46 |
| ResNeXt101+Faster RCNN+FPN | 69.53 | 63.91 |
| ResNeXt101+Cascade RCNN+FPN | 71.32 | 61.58 |

| | | |
|---|---|---|
| ResNet152+Cascade RCNN+FPN | 71.06 | 62.55 |

**[0136]** (2) Selection of a training sample data set for the human detection model 41

**[0137]** In the embodiments of the present application, for human detection in the video to be detected in a crowded scene, the human instance in the image is detected by a pre-trained human detection model 41. The human training sample set for training the human detection model 41 includes a human bounding box annotated with human instances, where each human instance is annotated with a whole body bounding box of the human body, and the human training sample set contains many images in the crowded scene. The data in the human training sample set is randomly scattered, and the sample set is divided into a training set, a verification set and a test set according to a specific proportion. Specifically, the human detection model 41 is trained using the video of the human instance in the training set and the data of the bounding box of the human instance to obtain a plurality of human detection intermediate models based on the human detection model 41. Each human detection intermediate model is verified by using the video verification set, and the human detection intermediate model with the best verification result is extracted. The human detection intermediate model is tested by using the video test set (preventing over-fitting). When the test result matches the verification result (the difference between the test result and the verification result is less than a preset threshold), the human detection intermediate model is output as the human detection model 41.

**[0138]** The selection and verification results of the training sample set for training the human detection model 41 are evaluated by using table 4 below.

**[0139]** Table 4 shows the performance results obtained by training the human detection model 41 through different training sample sets and performing human detection on the HIE verification set. Among them, the human detection model 41 is the baseline model. where the baseline model is a human detection model 41 with the Faster RCNN framework of ResNet50. The "+" in the table represents the data sets included in the training sample set. For example, HIE+CityPerson represents that the training sample set for training the baseline model includes the HIE data set and a CityPerson data set. All extra data represents a combined data set of a

COCO data set, the CityPerson data set, a CrownHuman data set, and a self-collected data set.

**[0140]** As can be seen from Table 4, the training sample set based on the HIE data set: the addition of the CrownHuman data set, COCO data set and CityPerson data set has the detection performance increased by 16.85 AP and reduced by 15.38 MMR. By adding the CrownHuman data set and the self-collected data set, the detection performance has the AP increased by 19.39 and the MMR reduced by 18.43. By adding the all extra data set, the detection performance has the AP increased by 19.68 and the MMR reduced by 18.84. It can be understood that the baseline model trained by the training sample set consisting of the HIE data set and the all extra data set has the best human detection performance, but compared to the baseline model trained by the training sample set consisting of the HIE data set, the CrowndHuman data set and the self-collected data set, the human detection performance has the AP increased by 0.33 and the MMR reduced by 0.41. The performance is not significantly improved, but the training time is three times higher due to the combination of the COCO data set. It can be understood that, in order to make a trade-off between the detection performance and the training time, the training sample set for training the human detection model 41 in the embodiments of the present application includes the HIE data set, the CrowndHuman data set, and the self-collected data set.

Table 4

| Training sample set for human detection | AP(%) | MMR(%) |
|---|---|---|
| HIE data set | 61.68 | 74.01 |
| HIE data set+COCO data set | 65.83 | 69.75 |
| HIE data set+CityPerson data set | 63.71 | 67.43 |
| HIE data set+CrowdHuman data set | 78.22 | 58.33 |
| HIE data set+self-collected data set | 69.39 | 60.82 |
| HIE data set+COCO data set+Crowd Human data set+CityPerson data set | 78.53 | 58.63 |
| HIE data set+Crowd Human data set+self-collected data set | 81.07 | 55.58 |

| | | |
|---|---|---|
| HIE data set+all extra data set | 81.36 | 55.17 |

**[0141]** (3) Training details of the human detection model 41

**[0142]** The detection method in the prior art predicts a single proposal box for a single human instance. Since the human instances in the detection scene overlap each other, some of the human instances are likely to be eliminated by the set non-maximum suppression (NMS) in the model detection process of a plurality of overlapped human instances, and finally only the bounding boxes of the human bodies corresponding to the remaining human instances are obtained. The NMS is used for filter out the human bounding box with a high Intersection over Union (IOU) in the detection results of the human detection model 41. The present application uses a method to predict human instances in the crowded scene, called "Crowd-Det". The key idea of Crowd-Det is to use the bounding box of the human body generated by the model as a proposal box, where each proposal box generated by the model does not recognize a single instance, but predicts a proposal box by taking the highly overlapped human instances as a whole, and then post-processing is performed on a human instance set contained in the proposal box, and finally a single human instance is obtained to obtain the bounding box of a single human body.

**[0143]** Specifically, the Crowd-Det algorithm above includes at least one of the following implementations, and it is described by taking a proposal box including two or more human instances, and a proposal box including human instance A or human instance B as an example:

a. EMD loss, used to determine human instance A and human instance B respectively corresponding to which of the bounding boxes of the two human instances generated by the proposal box through the convolutional neural network. Specifically, for example, the proposal box generates the bounding boxes of two human instances through the convolutional neural network, i.e., bounding box 1 and bounding box 2. Assuming that the bounding box 1 is the bounding box of the human instance A, and the bounding box 2 is the bounding box of the human instance B, the EMD loss is calculated. Assuming that the bounding box 2 is the bounding box of the human instance A, and the bounding box 1 is the bounding

box of the human instance B, the EMD loss is calculated. The bounding box corresponding to the human instance A with the smallest EMD loss value and the bounding box corresponding to the human instance B are used as the output result.

b. Refine module, configured to predict the bounding box for the second time. Specifically, in the embodiments of the present application, the bounding box of the human instance A and the bounding box of the human instance B obtained by the Crowd-Det algorithm through the EMD loss mechanism are input into the refine module together with a prediction box generated by the model, and the refine module generates the bounding boxes of two human bodies. The bounding boxes of the human bodies respectively corresponding to the human instance A and the human instance B are determined by using the EMD loss.

c. Set NMS, configured to insert an extra judgment to check whether the two boxes come from the same proposal box every time before the bounding box of a human body inhibits the bounding box of another human body in the NMS mechanism, and if yes, skip the inhibition. It can be understood that the bounding box of a human body inhibiting the bounding box of another human body in the NMS mechanism indicates that in the bounding box of the human body corresponding to two human bodies that overlap (for example, a higher degree of overlap), the bounding box of the human body in the front inhibits the bounding box of the occluded human body at the rear, and as a result, the bounding box of the occluded human body at the rear is eliminated. For example, in order to prevent the bounding box of the human body predicted by the overlapping human instance A or human instance B from being eliminated by the NMS, the original NMS is simply improved to obtain the foregoing "Set NMS" mechanism, that is, every time when the bounding box of a human body overlaps the bounding box of another human body, determining whether the bounding boxes of the two human bodies come from the same proposal box; if yes, skip the NMS inhibition.

| Table 5 | | |
|---|---|---|
| Method or model | AP(%) | MMR(%) |
| ResNet50+Faster RCNN | 81.36 | 55.17 |
| ResNet50+Faster RCNN +emd loss | 81.73 | 53.2 |
| ResNet50+Faster RCNN+ emd loss+refine module | 81.96 | 50.85 |
| ResNet50+Faster RCNN+ emd loss + refine module +set NMS | 82.05 | 49.63 |

**[0144]** Table 5 shows the verification result that the human detection model 41 is the baseline model, the baseline model is trained on the training sample set, and the human detection test is performed on the HIEvalidation set. The baseline model is a human detection model 41 with the Faster RCNN framework of ResNet50. The training sample set includes a combined data set of the HIE data set, the CityPerson data set, the COCO data set, the CrowndHuman data set, and the self-collected data set.

**[0145]** It can be seen from Table 5 that by adding the EMD loss of Crowd-Det to the baseline model, the detection performance has the AP increased by 0.37 and the MMR reduced by 1.97. By adding the emd loss processing and refine module processing on the baseline model,, the detection performance has the AP increased by 0.6 and the MMR reduced by 4.32. By adding the emd loss processing, refine module processing and set NMS processing to the baseline model, the detection performance has the AP increased by 0.69 and the MMR reduced by 5.54. It can be understood that, according to the results in the table, the Crowd-Det algorithm does improve the performance of crowd detection.

**[0146]** (4) KD regularization of the human detection model 41

**[0147]** In the embodiments of the present application, in the process of performing human detection on each image frame by the human detection model 41, the output layer of the human detection model 41 outputs the proposal box corresponding to the human body in the image, the possible category of the human body in the proposal box, and the probability corresponding to the category. The human detection model 41 selects the bounding box with the highest

category probability of the human body. In order to reduce the overfitting of the human detection model 41 to the human body in the detection process, In the training process of the human detection model 41, knowledge distillation (KD) regularization is additionally added. Specifically, in the model training of the human detection model 41, in addition to the conventional cross entropy (the first item of the following formula) as the loss function, KD regularization will provide an additional regularization formula (the second term of the formula below). For example, in training, As shown in Figure 11,for the human instance in the figure, the possible output of the human detection model 41 output layer and the probability distribution of the category are the probability of human being 0.5, the probability of tree 0.4, and the probability of chair 0.1. Normal training will only talk about the cross entropy of the probability distribution and a one-hot label as the loss function. KD regularization will also provide the probability distribution of a virtual teacher model and the KL divergence of the probability distribution of the model as an additional loss Function, performing KD regularization can make the model not excessively rely on pre-labeled labels when outputting the category and probability distribution in the result output layer, so that the output category and probability distribution of the human detection model 41 are more accurate.

**[0148]** The mathematical formula of KD regularization provided for the present application is as follows:

$$L_{reg} = (1-\alpha)H(q,p) + \alpha D_{KL}(p^d_\tau, p_\tau)$$

where H represents the cross entropy, p represents the probability of category output by the detection model, q represents the ground truth of the category, $D_{KL}$ represents KL divergence, $p_\tau$ represents the probability distribution of the category output by the detection model after temperature t soften, $p^d{}_\tau$ represents the artificially designed probability distribution, and the distribution formula is as follows:

$$p^d(k)=\begin{cases} a & \text{if } k = c, \\ (1-a)/(K-1) & \text{if } k \neq c, \end{cases}$$

**[0149]** Finally, based on the analysis abovethe HIE data set, the crowdhuman data set and

the self-collected data set are adopted to train two human detection models 41, annotated as human detection model 1 and human detection model 2. Detection performance verified on HIE validation set: The backbone network ResNet152 of the human detection model 1 and the framework selection cascade RCNN are combined with FPN, and Crowd-Det and KD regularization post-processing are performed on the bounding box of the human body generated by the output layer of the model, and the detection performance is 83.21 AP. The backbone network ResNeXt101 of the human detection model 2 and the framework selection cascade RCNN are combined with FPN, and Crowd-Det and KD regularization post-processing are performed on the bounding box of the human body generated by the output layer of the model, and the detection performance is 83.78 AP. A detection model obtained by fusing the human detection model 1 and the human detection model 2 in a weight ratio of 1:1 is taken as the human detection model 41 in the embodiments of the present application.

**[0150]** In the action recognition solution of the present application, the human pose estimation model 44 is used to recognize the poses of the human instances in the human bounding box, generate human joint points, and estimate the human poses based on the relative positions of the joint points. Specifically, the function and training process of the model are as follows.

**[0151]** Training purpose of the human pose estimation model 44

**[0152]** In the embodiments of the present application, the human pose estimation model 44 is trained based on the training sample set of pose estimation, and the detection performance of the human pose model is determined by the evaluation indicator m-AP. m-AP reflects the accuracy and recall rate of human body posture, mAP is mean Average Precision. The larger the value of mAP, the more accurate the detection performance of the human posture estimation model 44.In the embodiments of the present application, Table 6 shows the test results of the human pose estimation model 44 in different scenes. It can be seen from Table 6 that remarkable performance has been achieved in both conventional videos and high-resolution videos. For example, the AP of the conventional video "hm_in_passage" and the AP of high-resolution

video "hm_in_crossroad" are 90.14% and 75.76%, respectively. The effect is poor in the video with crowded scenes and low quality, for example, in the super crowded video "hm_in_dining_room2", the AP is 25.94%.

| Table 6 | |
|---|---|
| Video name | Test result AP (%) |
| hm_in_waiting_hall | 65.827 |
| hm_in_bus | 55.4518 |
| hm_in_dining_room2 | 25.9449 |
| hm_in_lab2 | 70.33 |
| hm_in_subway_station | 47.2997 |
| hm_in_passage | 90.1478 |
| hm_in_fighting | 59.8902 |
| hm_in_shopping_mall3 | 62.7075 |
| hm_in_restaurant | 67.4902 |
| hm_in_accident | 56.9365 |
| hm_in_stair3 | 49.1054 |
| hm_in_crossroad | 75.764 |
| hm_in_robbery | 52.2467 |
| weighted Average | 56.3375 |

**[0153]** (2) Selection of a training sample data set for the human pose estimation model 44

**[0154]** In the embodiments of the present application, the training sample set contains about 660,500 annotated poses. Considering the existence of large-scale similar data due to frame-by-frame annotation, it is necessary to collect additional data sets to improve performance. The extra training data set used comes from two aspects:

**[0155]** On the one hand, three mainstream pose estimation data sets, such as a COCO data set, an MPII data set and an AI Challenger data set, are integrated into training data. The COCO data set contains 250,000 human instances with 17 key points. The MPII data set consists of

25K images, including more than 40K human instances and 16 body joints with annotations. The AI Challenger data set consists of about 700K human instances and 14 body joints with annotations. Since the annotated key points in these data sets do not completely overlap with official labels, for each data set, in the embodiments of the present application, the overlapping key points are used for training.

**[0156]** On the other hand, the self-collected data sets with similar scenes are merged into the training set. There are no more than 30K poses of the self-collected data sets, which are far less than the training data of the HIE data sets.

**[0157]** In the embodiments of the present application, the pose training sample set is used to train the human pose model. Specifically, the height or width of the bounding box of the human body (of the pose training sample set) is expanded to a fixed 4:3 aspect ratio, and then the box is cut from the image and adjusted to a fixed size of 256×192 or 384×288. Data augmentation includes random rotation ([-45°, 45°]), random scale (0.65, 1.35) and flip. Half-body data enhancement is also used to train the model. The half-body data enhancement refers to pre-dividing the joint points of the human instance in the posture training sample into upper body joint points and lower body joint points, and then randomly selects the joint points of the human binstance during model training Or the upper half joint points of the human body instance or the lower half joint points of the human instance to train the human posture estimation model 44.(3) Description of test details of the human pose estimation model 44

**[0158]** In the embodiments of the present application, taking the video of the crowded scene as an example, on the one hand, multi-scale evaluation is adopted in the test process. Specifically, the bounding box of the human body is rescaled to obtain the bounding boxes of new human bodies with different proportions, then the bounding boxes are cut to the original size and flipped to obtain the flipped corresponding boxes. The generated boxes are fed into the network to generate heat maps. These heat maps are averaged, and the highest response is searched to obtain the positions of key points. The scale factors used are 0.7, 1.0 and 1.3. In addition, in complex and crowded scenes, the test is easily subjected to redundant and incorrect

bounding boxes of the human body. Pose NMS is used to eliminate redundancy with similar and low confidence levels. On the other hand, multi-scale input is adopted in the testing process, and "multi-scale input" means implementation by changing Input image length and width. The size is entered and the results are fused during the test to train a plurality of groups of parameters. If "multi-scale input" is adopted, the results are obtained by fusing the input sizes of 256×192 and 384×288; and if "multi-scale input" is not used, the input size is only 256×192.

**[0159]** Table 7 shows different composition experiment performance indicators of pose estimation in the training test process of the present application, where the test results are validated by human pose estimation test in the video of crowded scenes. It can be seen from Table 7 that the pose estimation model uses the HRNet network to train the pose estimation model based on the HIE data set, and the performance indicator obtained by test verification is 52.45%. The pose estimation model uses the HRNet network and is trained based on the HIE data set. Multi-scale evaluation processing is added in the test process, and the performance indicator obtained by test verification is 52.9%. The pose estimation model uses the HRNet network and is trained based on the pose training sample set. The performance indicator obtained by test verification is 53.82%. The pose estimation model fuses the HRNet network with the SimpleNet network to perform human pose estimation, and the pose estimation model is trained based on the pose training sample set. Multi-scale evaluation processing is added in the test process, and the performance indicator obtained by test verification is 55.52%. The pose estimation model fuses the HRNet network with the SimpleNet network to perform human pose estimation, and the pose estimation model is trained based on the pose training sample set. Multi-scale evaluation and multi-scale input processing are added in the test process, and the performance indicator obtained by test verification is 56.04%. The pose estimation model fuses the HRNet network with the SimpleNet network to perform human pose estimation, and the pose estimation model is trained based on the pose training sample set. Multi-scale evaluation and multi-scale input processing are added in the test process. The generated joint points are subjected to optical flow smoothing and the performance indicator obtained by test verification

is 56.34%. The pose training sample sets include HIE data sets, COCO data sets, MPII data sets, AI Challenger data sets and self-collected data sets.

| Table 7 | |
|---|---|
| Composition of a pose estimation method | Performance indicator W-AP (%) |
| HRNet | 52.45 |
| HRNet+ multi-scale evaluation | 52.9 |
| HRNet+Extra Data | 53.82 |
| HRNet+SimpleNet+ multi-scale evaluation +Extra Data | 55.52 |
| HRNet+SimpleNet+ multi-scale evaluation + multi-scale input +Extra Data | 56.04 |
| HRNet+SimpleNet+ multi-scale evaluation + multi-scale input + optical flow smoothing | 56.34 |

**[0160]** FIG. 12 shows a structural block diagram of a video-based human pose estimation apparatus according to some embodiments of the present application. As shown in FIG. 12, a video-based human pose estimation apparatus 1200 includes:

an image acquisition module (1202), configured to obtain a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances; a human detection module (1204), configured to generate initial human bounding boxes of the human instances in each multi-human image frame; a human correction module (1206), configured to correct the initial human bounding boxes 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 detected human bounding boxes, where the position difference data includes a position difference of the human instances in the multi-human image between the previous image frame and the multi-human image; and a pose estimation module (1208), configured to recognize poses of the human instances in the multi-human image based on the detected human bounding boxes of the multi-human image.

**[0161]** It can be understood that the video-based human pose estimation apparatus 1200 shown in FIG. 12 corresponds to the video-based human pose estimation method provided in the present application. The technical details in the above detailed description of the video-based human pose estimation method provided in the present application are still applicable to the video-based human pose estimation apparatus 1200 shown in FIG. 12. Refer to the above for detailed description, and it will not be repeated here.

**[0162]** FIG. 13 shows a structural block diagram of a video-based human pose estimation apparatus according to some embodiments of the present application. As shown in FIG. 13, a video-based human pose estimation apparatus 1300 includes:

an image acquisition module (1302), configured to obtain a plurality of multi-human image frames in a video, where each multi-human image frame includes a plurality of human instances;

a human detection module (1304), configured to generate detected human bounding boxes of the human instances in each multi-human image frame;

a pose detection module (1306), 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;

a pose correction module (1308), 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, where the position difference data between the multi-human image and the previous image frame includes 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 includes a position difference of the human instances in the multi-human image between the multi-human image and the next image frame; anda pose estimation module (1310),

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.

**[0163]** It can be understood that the video-based human pose estimation apparatus 1300 shown in FIG. 13 corresponds to the video-based human pose estimation method provided in the present application. The above technical details in the specific description of the video-based human pose estimation method provided in the present application are still applicable to the video-based human pose estimation apparatus 1300 shown in FIG. 13. Reference may be made to the specific description above, and details are not described herein again.

**[0164]** FIG. 14 shows a block diagram of an electronic device according to an embodiment of the present application. The electronic device may be the foregoing video processing device 300. FIG. 14 schematically illustrates an exemplary electronic device 1400 according to a plurality of embodiments. In one embodiment, the electronic device 1400 may include one or more processors 1404, a system control logic 1408 connected to at least one of the processors 1404, a system memory 1412 connected to the system control logic 1408, a non-volatile memory (NVM) 1416 connected to the system control logic 1408, and a network interface 1420 connected to the system control logic 1408.

**[0165]** In some embodiments, the processor 1404 may include one or more single-core or multi-core processors. In some embodiments, the processor 1404 may include any combination of a general-purpose processor and a special-purpose processor (for example, a graphics processor, an application processor, a baseband processor, etc.). In an embodiment where the electronic device 1400 adopts an evolved Node B (eNB) or radio access network (RAN) controller, the processor 1404 may be configured to execute conforming embodiments, e.g., one or more of the plurality of embodiments shown in FIG. 3. For example, the processor 1404 can be used to perform the video-based human pose estimation method, such as generating human bounding boxes and joint points of human instances in a plurality of multi-human image frames, and obtaining human bounding boxes and joint points based on correction by optical flow between a plurality of adjacent multi-human image frames.

**[0166]** In some embodiments, the system control logic 1408 may include any suitable interface controller to provide any suitable interface to at least one of the processors 1404 and/or any suitable device or component in communication with the system control logic 1408.

**[0167]** In some embodiments, the system control logic 1408 may include one or more memory controllers to provide an interface connected to the system memory 1412. The system memory 1412 can be configured to load and store data and/or instructions. In some embodiments, the memory 1412 of the system 1400 may include any suitable volatile memory, such as a suitable dynamic random access memory (DRAM).

**[0168]** The NVM/memory 1416 may include one or more tangible, non-transient computer readable media for storing data and/or instructions. In some embodiments, the NVM/memory 1416 may include any suitable non-volatile memory such as a flash memory and/or any suitable non-volatile storage device, such as at least one of a hard disk drive (HDD), a compact disc (CD) drive, or a digital versatile disc (DVD) drive.

**[0169]** The NVM/memory 1416 may include a part of the storage resources on an apparatus where the electronic device 1400 is installed, or it may be accessed by the device, but not necessarily a part of the device. For example, the NVM/memory 1416 can be accessed through the network via the network interface 1420.

**[0170]** In particular, the system memory 1412 and the NVM/memory 1416 may respectively include: a temporary copy and a permanent copy of the instruction 1424. The instruction 1424 may include an instruction that, when executed by at least one of the processors 1404, causes the electronic device 1400 to implement the method shown in FIG. 1. In some embodiments, the instruction 1424, hardware, firmware, and/or software components thereof may additionally/alternatively be placed in the system control logic 1408, the network interface 1420, and/or the processor 1404.

**[0171]** The network interface 1420 may include a transceiver, configured to provide a radio interface for the electronic device 1400, and then communicate with any other suitable device (such as a front-end module and an antenna) through one or more networks. In some

embodiments, the network interface 1420 may be integrated in other components of the electronic device 1400. For example, the network interface 1420 may be integrated in at least one of the processor 1404, the system memory 1412, the NVM/memory 1416, or a firmware device (not shown) with instructions. The electronic device 1400 implements the method shown in FIG. 5 when at least one of the processors 1404 executes the instruction.

**[0172]** The network interface 1420 may further include any suitable hardware and/or firmware to provide a multi-input multi-output radio interface. For example, the network interface 1420 may be a network adapter, a wireless network adapter, a telephone modem and/or a wireless modem.

**[0173]** In one embodiment, at least one of the processors 1404 may be packaged together with the logic of one or more controllers for the system control logic 1408 to form a system in package (SiP). In one embodiment, at least one of the processors 1404 may be integrated on the same die with the logic of one or more controllers for the system control logic 1408 to form a system on chip (SoC).

**[0174]** The electronic device 1400 may further include an input/output (I/O) device 1432. The I/O device 1432 may include a user interface which enables a user to interact with the electronic device 1400. The design of a peripheral component interface enables a peripheral component to interact with the electronic device 1400. In some embodiments, the electronic device 1400 further includes a sensor, configured to determine at least one of environment conditions and position information related to the electronic device 1400.

**[0175]** In some embodiments, the user interface may include, but is not limited to, a display (e.g., a liquid crystal display, a touchscreen display, etc.), a loudspeaker, a microphone, one or more cameras (e.g., a still image camera and/or a camcorder), a flashlight (e.g., an LED flash), and a keyboard.

**[0176]** In some embodiments, the peripheral component interface may include, but is not limited to, a non-volatile memory port, an audio jack, and a power interface.

**[0177]** In some embodiments, the sensor may include, but is not limited to, a gyroscope

sensor, an accelerometer, a proximity sensor, an ambient light sensor, and a positioning unit. The positioning unit may also be a part of the network interface 1420 or interact with the network interface 1420 to communicate with components of the positioning network (e.g., a global positioning system (GPS) satellite).

**[0178]** According to the embodiments of the present application, FIG. 15 shows a block diagram of a SoC 1500. In FIG. 15, similar components have the same reference numerals. In addition, the dashed-line box is an optional feature of more advanced SoCs. In FIG. 15, the SoC 1500 includes: an interconnection unit 1550, coupled to an application processor 1510; a system agent unit 1570; a bus controller unit 1580; an integrated memory controller unit 1540; a group of or one or more coprocessors 1520 which may include an integrated graphics logic, an image processor, an audio processor, and a video processor; a static random access memory (SRAM) unit 1530; and a direct memory access (DMA) unit 1560. In one embodiment, the coprocessor 1520 includes a special-purpose processor, such as a network or communication processor, a compression engine, a GPGPU, a high-throughput MIC processor, or an embedded processor.

**[0179]** Embodiments of the mechanism disclosed in the present application can be implemented in hardware, software, firmware, or a combination of these implementation methods. The embodiments of the present application can be implemented as a computer program or program code executed on a programmable system including at least one processor and a storage system (including volatile and non-volatile memories and/or storage elements), at least one input device and at least one output device.

**[0180]** Program code can be applied to an input instruction to perform the functions described in the present application and generate output information. The output information can be applied to one or more output devices in a known manner. For the purpose of the present application, a processing system includes any system having a processor, such as, a digital signal processor (DSP), a microcontroller, an application-specific integrated circuit (ASIC), or a microprocessor.

**[0181]** The program code can be implemented in an advanced programming language or an

object-oriented programming language to communicate with the processing system. When needed, assembly language or machine language can also be used to implement the program code. In fact, the mechanism described in the present application is not limited to the scope of any particular programming language. In either case, the language can be a compiled language or an interpreted language.

**[0182]** In some cases, the disclosed embodiments may be implemented in hardware, firmware, software, or any combination thereof. The disclosed embodiments can also be implemented as instructions carried by or stored on one or more transient or non-transient machine-readable (e.g., computer-readable) storage media, and the instructions can be read and executed by one or more processors. For example, the instructions can be distributed through a network or other computer-readable media. Therefore, the machine-readable media may include any mechanisms for storing or transmitting information in a machine (e.g., computer) readable form, including, but not limited to, a floppy disk, an optical disc, a compact disc read-only memory (CD-ROM), a magneto-optical disc, a read-only memory (ROM), a random access memory (RAM), an erasable programmable read-only memory (EPROM), an electrically erasable programmable read-only memory (EEPROM), a magnetic or optical card, a flash memory, or a tangible machine-readable memory used to transmit information (for example, carriers, infrared signals, digital signals, etc.) using the Internet with electrical, optical, acoustic, or other forms of propagating signals. Therefore, the machine-readable medium includes any type of machine-readable medium suitable for storing or transmitting electronic instructions or information in a machine (e.g., computer) readable form.

**[0183]** In the accompanying drawings, some structural or method features may be shown in a specific arrangement and/or order. However, it should be understood that such a specific arrangement and/or order may not be required. Instead, in some embodiments, these features may be arranged in a different manner and/or order than that shown in the illustrative drawings. In addition, the inclusion of structural or method features in a particular figure does not imply that such features are required in all embodiments, and in some embodiments, these features

may not be included or may be combined with other features.

**[0184]** It should be noted that each unit/module mentioned in each device embodiment of the present application is a logical unit/module. Physically, a logical unit/module can be a physical unit/module or a part of the physical unit/module, and can also be implemented by a combination of a plurality of physical units/modules. The physical implementation of these logical units/modules is not the most important. The combination of the functions implemented by these logical units/modules is the key to solve the technical problems raised in the present application. In addition, in order to highlight the innovative part of the present application, the foregoing device embodiments of the present application do not introduce units/modules that are not closely related to solving the technical problems raised in the present application, which does not indicate that the foregoing device embodiments do not have other units/modules.

**[0185]** It should be noted that in the examples and specification of the present patent, relational terms such as first and second are only used to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply any such actual relationship or sequence between these entities or operations. Moreover, the term "comprise", "include", or any other variants thereof is intended to encompass a non-exclusive inclusion, so that the process, method, article, or device including a series of elements not only includes those elements, but also includes other elements that are not explicitly listed, or elements that are inherent to such a process, method, article, or device. An element defined by the phrase "including a ..." does not exclude the presence of the same element in the process, method, article, or device that includes the element, without more restrictions.

**[0186]** Although the present application has been illustrated and described by referring to some preferred embodiments of the present application, those of ordinary skill in the art should understand that various changes can be made in form and details without departing from the spirit and scope of the present application.

## 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 image frame comprises a plurality of human instances;

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 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 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-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:

$$\hat{\mathrm{I}}_{\mathrm{i}}^{\mathrm{t}}=\mathrm{I}_{\mathrm{i}}^{\mathrm{t}-1}+\mathrm{F}^{\mathrm{t}-1\rightarrow\mathrm{t}},$$

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

wherein $\mathrm{I}_{\mathrm{i}}^{\mathrm{t}-1}$ 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, $\mathrm{F}^{\mathrm{t}-1\rightarrow\mathrm{t}}$ represents the position difference data between the (t–1)-th image frame and the t-th image frame, $\hat{\mathrm{I}}_{\mathrm{i}}^{\mathrm{t}}$ represents a coordinate point set of the i-th predicted human bounding box in the t-th image

frame obtained based on $F^{t-1\to t}$, $I_i^t$ 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 $\alpha$ represents a weight coefficient synthesized by $I_i^t$ and $\hat{I}_i^t$, 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:

$$\hat{J'}_i^t = J_i^{t-1} + F^{t-1\to t},$$

$$\widehat{J''}_i^{t}=J_i^{t+1}-F^{t\to t+1},$$

$$\tilde{J}_i^{t}=\beta\cdot\widehat{J'}_i^{t}+\beta\cdot\widehat{J''}_i^{t}+(1-2\beta)\cdot J_i^{t},$$

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, $F^{t-1\to t}$ represents position difference data between the (t–1)-th image frame and the t-th image frame, $\widehat{J'}_i^{t}$ 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 $F^{t-1\to t}$; $J_i^{t+1}$ 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; $F^{t\to t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\widehat{J''}_i^{t}$ 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 $F^{t\to t+1}$; $\tilde{J}_i^{t}$ 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 $\beta$ represents a weight coefficient synthesized between $J_i^{t}$, $\widehat{J''}_i^{t}$ and $\widehat{J'}_i^{t}$, 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:

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 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.

[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 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 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;

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; 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:

$$\widehat{J'}_i^{\mathrm{t}}=J_i^{\mathrm{t}-1}+F^{\mathrm{t}-1\rightarrow\mathrm{t}},$$

$$\widehat{J''}_i^{\mathrm{t}}=J_i^{\mathrm{t}+1}-F^{\mathrm{t}\rightarrow\mathrm{t}+1},$$

$$\tilde{J}_i^t=\beta\cdot\widehat{J'}_i^{\mathrm{t}}+\beta\cdot\widehat{J''}_i^{\mathrm{t}}+(1-2\beta)\cdot J_i^{\mathrm{t}},$$

wherein $J_i^{\mathrm{t}-1}$ 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, $F^{t-1\to t}$ represents position difference data between the (t–1)-th image frame and the t-th image frame, $\widehat{J'}_i^t$ 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 $F^{t-1\to t}$; $J_i^{t+1}$ 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; $F^{t\to t+1}$ represents position difference data between the t-th image frame and the (t+1)-th image frame; $\widehat{J''}_i^t$ 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 $F^{t\to t+1}$; $\tilde{J}_i^t$ 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 $J_i^t$, $\widehat{J''}_i^t$ and $\widehat{J'}_i^t$, wherein i and t are positive integers.

[Claim 13] The method according to claim 11 or 12, wherein the generating detected 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-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.

[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 multi-human image frames by using the following formulas to obtain the i-th detected human bounding box:

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

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

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, $F^{t-1\to t}$ 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 $F^{t-1\to t}$, $I_i^t$ 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 $\alpha$ 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 instances;

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

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; and

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:

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;

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

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;

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; and

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