

TITLE OF INVENTION: VIDEO-BASED HUMAN ACTION RECOGNITION METHOD AND APPARATUS, MEDIUM AND ELECTRONIC DEVICE

TECHNICAL FIELD

5 **[0001]** The present application relates to the field of computer vision, and in particular, to a video-based human action recognition method and apparatus, a medium and an electronic device.

BACKGROUND

10 **[0002]** In many scenes, it is necessary to recognize human actions in videos to determine the category of the human actions. For example, in a surveillance system, in order to perceive the intention of a person, it is necessary to recognize human actions in videos obtained by the surveillance system. In a human-computer interaction system, in order to understand human behavior, it is necessary to recognize the human actions in the videos obtained by the human-computer interaction system.

15 **[0003]** In order to recognize human actions in the videos, the existing solution is to extract some frames of images from a video to be processed, then use a convolutional neural network to perform feature extraction on the images, and then determine the action category of the human and the confidence level of each action category according to a feature vector of a region where the human body extracted from the image is located, and then determine the action
20 category with the highest confidence level (moreover, the confidence level is greater than a preset threshold) as the action category of the human in the video.

[0004] The existing solution directly extracts the features of a part of the region in the image to perform action recognition, but the direct and simple extraction of the features of a part of the region in the image may not well reflect the human body's action features in the video frame

scene, resulting in a poor action recognition effect and low accuracy.

SUMMARY

[0005] Embodiments of the present application provide a video-based human action recognition method, which can perform post-processing on the action category and probability of the human instances according to the ambient environment features in video frames of the human instances in a video clip, so that the accuracy of the finally predicted action category of the human instances is higher.

[0006] In a first aspect, the embodiments of the present application disclose a video-based human action recognition 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 human instances in each multi-human image frame; determining an image scene of each multi-human image frame, one image scene corresponding to one or more categories of actions; and recognizing the actions of human instances in the plurality of multi-human image frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames. 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, in the process of recognizing the actions of the human instances in the plurality of multi-human image frames, consecutive frames (such as 40 frames) of multi-human images can be processed at a time, that is, one video clip in the video can be processed at a time. Moreover, one image scene (i.e., a scene category below) corresponds to one or more categories of actions, that is, the scene category below is coupled to the action category, and when the human instance is in a specific scene, it is usually in a

corresponding specific action, that is, the image scene where the human instance is located can reflect the action of the human instance. For example, when the human instance is in the image scene "stairs", the action of the human instance is usually "walking up and down stairs". In this way, during recognition of the action of the human instance in the detected human bounding box of the plurality of multi-human image frames, with reference to the image scene where the human instance is located, that is, with reference to the relationship between the human instance and the scene where the human instance is located, the accuracy of the action category of the finally recognized human instance can be improved.

[0007] It should be noted that in the embodiments of the present application, for ease of description, different names are used in different locations to describe images involved in human action recognition, such as "video frames" and "multi-human images", with the essence unchanged.

[0008] In a possible implementation of the first aspect, the recognizing the actions of human instances in the plurality of multi-human image frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames includes: recognizing the actions of the human instances in the plurality of multi-human image frames based on the detected human bounding boxes of the plurality of multi-human image frames, to obtain an initial probability of each of a plurality of predefined actions corresponding to the human instance; correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the image scenes of the plurality of multi-human image frames, to obtain a corrected probability of each predefined action; and taking the predefined action with the highest probability in each multi-human image frame as the action recognized by the human instance. It can be understood that the initial probability of each of the predefined actions corresponding to the human instance can be determined by a neural network model used for human action recognition (i.e., the action detection model below). The plurality of predefined actions are a plurality of action classifications provided by the action detection model. However, the action detection model may be a conventional or pre-trained network, which is not specifically limited in the embodiments of the present application.

Specifically, the action detection model can determine the action category and initial probability distribution of the human instance by performing action recognition on the human instance. The correcting the initial probability distribution of each action category of the human instance is specifically adjusting the weight of each probability, and normalizing all the adjusted weights to obtain the action with the highest corrected probability of the human instance. Therefore, weighted processing (i.e., post-processing below) is performed again on the probability of the action category of the human instance predicted by the action detection model based on the image scenes of the plurality of multi-human image frames, to correct the initial probability distribution of the action category of the human instance detected by the action detection model, so that the accuracy of the finally detected action category of the human instance is higher.

[0009] In a possible implementation of the first aspect, the correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the image scenes of the plurality of multi-human image frames, to obtain a corrected probability of each predefined action includes: correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames and the image scenes of the plurality of multi-human image frames, to obtain the corrected probability of each predefined action. It can be understood that the degree of overlap between adjacent detected human bounding boxes of each multi-human image frame can reflect the degree of overlap between adjacent human instances in each multi-human image frame. However, the degree of overlap between adjacent human instances can reflect the actions of these human instances. For example, when the degree of overlap between adjacent human instances is higher, the actions of these human bodies are usually multi-person actions, such as running together. On the contrary, when the degree of overlap between adjacent human instances is lower, the actions of these human bodies are usually single-person actions, such as single-person running. In this case, weighted processing is performed again on the probability of the action category of the human instance predicted by the action detection model based on the degree of overlap between adjacent detected human bounding boxes of the plurality of

multi-human image frames, to correct the initial probability of the action category of the human instance detected by the action detection model, so that the accuracy of the finally detected action category of the human instance is higher.

[0010] In a possible implementation of the first aspect, the degree of overlap between adjacent detected human bounding boxes is represented by an Intersection over Union (IoU) between adjacent detected human bounding boxes. It can be understood that the higher the Intersection over Union(IoU) between the adjacent detected human bounding boxes is, the higher the degree of overlap between the adjacent detected human bounding boxes is, and vice versa.

[0011] In a possible implementation of the first aspect, the correcting the initial probability of each of the predefined actions corresponding to the human instance based on the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames includes: when the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames is greater than a predetermined threshold, increasing a probability of a predefined action related to a plurality of persons among the plurality of predefined actions corresponding to the human instances; or when the degree of overlap between adjacent detected human bounding boxes of the plurality of multi-human image frames is less than or equal to the predetermined threshold, increasing a probability of a predefined action related to a single person among the plurality of predefined actions corresponding to the human instances. The value of the foregoing predetermined threshold may be determined according to actual needs, such as 5, which is not specifically limited in the embodiments of the present application. It can be understood that when the degree of overlap between the adjacent detected human bounding boxes in the plurality of multi-human image frames is greater than the predetermined threshold, it indicates that these human instances are more likely to produce multi-person actions. Therefore, the probability of multi-person related actions of these human instances can be increased, for example, the probability of the action (i.e., the action category) "running together" is increased; in this case, the probability of the single-person related action category in these human instances is reduced accordingly, for

example, the probability of the action "single-person running" is reduced. Similarly, when the degree of overlap between the adjacent detected human bounding boxes in the plurality of multi-human image frames is less than or equal to the predetermined threshold (such as 4), it indicates that these human instances are more likely to produce single-person actions. Therefore, the probability of single-person related actions of these human instances can be increased, for example, the probability of the "single-person running" is increased, and the probability of the multi-person related action category in these human instances is reduced accordingly, for example, the probability of the action "running together" is reduced.

[0012] In a possible implementation of the first aspect, the correcting the initial probability of each of the predefined multi-category actions corresponding to the human instance based on the image scenes of the plurality of multi-human image frames includes: for the human instance in a region where the image scenes are located in the plurality of multi-human image frames, increasing probabilities of the predefined multi-category actions corresponding to the human instance and the predefined action corresponding to the determined image scene based on the image scenes of the plurality of multi-human image frames. It can be understood that when the human instance is located in a region where the determined image scene is located in the plurality of multi-human image frames, it indicates that these human instances are more likely to produce actions corresponding to the image scene. Therefore, the probability of the action category of these human instances corresponding to the determined image scene can be increased. For example, when the human instances are located in the region where the image background "stairs" is located, the probability of the action "walking up and down stairs" of these human instances may be increased, and the probability of other actions (such as the action "playing football") of these human instances may be reduced accordingly.

[0013] In a possible implementation of the first 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. It can be understood that 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 running 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 action recognition based on the human bounding box.

[0014] 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, which is not limited in the embodiments of the present application.

[0015] 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 formula to obtain the i-th detected human bounding box:

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

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

[0016] 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 running status of a human instance in the (t-1)-th image frame reflected by $F^{t-1 \rightarrow t}$ 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.

5 **[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 model 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
10 plurality of networks above, the first neural network model has all the functions of these networks. Specifically, the first neural network model may be a human detection model below.

[0018] In a second aspect, the embodiments of the present application disclose a video-based human action recognition apparatus, including: an acquisition module, configured to obtain a plurality of multi-human image frames in a video, where each multi-human image
15 frame includes a plurality of human instances; a first generation module, configured to generate detected human bounding boxes of human instances in each multi-human image frame; a determining module, configured to determine an image scene of each multi-human image frame, one image scene corresponding to one or more categories of actions; and a recognition module, configured to recognize the actions of human instances in the plurality of multi-human image
20 frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames.

[0019] In a third 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
25 action recognition method according to the first aspect.

[0020] In a fourth 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 action recognition method according to the first aspect.

BRIEF DESCRIPTION OF DRAWINGS

[0021] FIG. 1 shows a schematic diagram of semantic segmentation results according to some embodiments of the present application;

[0022] FIG. 2 shows a schematic diagram of an optical flow between two consecutive frames according to some embodiments of the present application;

[0023] FIG. 3 shows a diagram of a human action recognition scene according to some embodiments of the present application;

[0024] FIG. 4 shows a schematic diagram of a composition architecture of an action recognition scene according to some embodiments of the present application;

[0025] FIG. 5 shows a flowchart of an action recognition method according to some embodiments of the present application;

[0026] FIG. 6 shows a schematic diagram of a video clip sequence according to some embodiments of the present application;

[0027] FIG. 7 shows a schematic diagram of a bounding box output by a human detection model according to some embodiments of the present application;

[0028] FIG. 8 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;

[0029] FIG. 9 shows a block diagram of a SlowFast network composition according to some embodiments of the present application;

[0030] FIG. 10 shows a schematic diagram of action recognition based on an AIA model according to some embodiments of the present application;

[0031] FIG. 11 shows a schematic structural diagram of a multi-interaction relationship modeling fusion network according to some embodiments of the present application;

[0032] FIG. 12a shows a visualized schematic diagram of a stair scene output by a semantic segmentation model according to some embodiments of the present application;

[0033] FIG. 12b shows a visualized schematic diagram of a passage scene output by a semantic segmentation model according to some embodiments of the present application;

[0034] FIG. 13 shows a schematic diagram of a coupling relationship established between a scene category and an action category according to some embodiments of the present application;

[0035] FIG. 14 shows a block diagram of possible output categories and probabilities when a human body detection model detects a human instance;

[0036] FIG. 15 shows a block diagram of an action recognition apparatus according to some embodiments of the present application;

[0037] FIG. 16 shows a block diagram of a system according to some embodiments of the present application;

[0038] FIG. 17 shows a block diagram of a system on chip (SoC) according to some embodiments of the present application.

DESCRIPTION OF EMBODIMENTS

[0039] Illustrative embodiments of the present application include, but are not limited to, a video-based human action recognition method and apparatus, a medium and an electronic device.

[0040] The video-based human action recognition method in the present application is suitable for various video processing scenes, especially video processing of crowded scenes.

For example, the scene is applied to robot vision processing, video surveillance, human action recognition, and animal action recognition, etc., and is specifically used for pedestrian action recognition on the road, dangerous action recognition in special scenes, etc.

[0041] The video-based human action recognition method disclosed in the present application can first extract behavior representation information of a human instance in a video frame through an action detection model, and then with reference to the position of the human instance in the video frame extracted by the human detection model, perform action recognition on the human instance to obtain the probability distribution of the action category of the human

instance. Then, post-processing is performed on the action category and probability of the human instance with reference to the ambient environment features of the human instance in the video clip, that is, weighted processing is performed again on the action category and probability of the human instance predicted by the action detection model. The ambient environment features include the scene category in the video frame, the IoU of the bounding box of the human instance in the video frame, etc. It can be understood that the action category of the human body is predicted with reference to the interaction relationship between the human instance and the ambient environment, and then the weight of the probability distribution of the action category is reallocated, so that the accuracy of the finally predicted action category of the human instance is higher.

[0042] 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:

[0043] Semantic segmentation (SS): it is an image processing method in computer vision, which can recognize images at the pixel level, perform intensive prediction on each pixel, annotate the object category to which each pixel in the image belongs, and divide the image into several categories, such as scenes and objects (including human bodies, animals, food, daily necessities, transportation means, etc.). It can be understood that if an image describes a scene of pedestrians walking up and down stairs, the semantic segmentation technology can classify the pixels representing people in the image into one category and the pixels representing the stair background into one category. For example, as shown in FIG. 1, FIG. 1(a) shows an image of pedestrians walking up and down stairs, FIG. 1(b) shows an image of a stair scene extracted through semantic segmentation, It can be seen from the figure that the scene information extracted through semantic segmentation can reflect the ambient environment information where pedestrians are located.

[0044] 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 is a person. For ease of description, both "human instance" and "human body" herein refer to a person, and the "bounding box of the human body" or "human bounding box" herein refers to the position of the human body in the image.

[0045] 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:

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

where $F^{t-1 \rightarrow t}$ 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. Δx and Δy represent the position differences from the (t-1)-th frame to the t-th frame.

[0046] For ease of description of the technical solution of the present application, description is made below by taking human action recognition under video surveillance as an example.

[0047] FIG. 3 shows human action recognition of a subway stair scene under video surveillance according to the embodiments of the present application. As shown in FIG. 3, the scene includes three cameras 301 and a video processing device 300, but it can be understood that the action recognition scene applicable to the technical solution of the present application may include any number of cameras, which is not limited to three.

[0048] The cameras 301 are configured to capture or shoot a video stream containing a human body, and the video processing device 300 is configured to perform action recognition on the human bodies in the obtained video stream containing the human body, such as walking up and down stairs, falling, fighting, and stealing.

[0049] 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.

[0050] 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.

[0051] As an example, in the scene of performing action recognition on the human body in video surveillance, the cameras 301 may be surveillance cameras disposed in places such as roads or subway entrances/exits for shooting video streams of pedestrians in these places. In practical applications, a larger number of cameras 301 may be included in the above-mentioned action recognition scene, for example, including cameras 301 disposed at various positions of a subway station.

[0052] In some embodiments, the camera 301 and the video processing device 300 are two 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 different components in the same mobile phone.

[0053] 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.

[0054] In some embodiments of the present application, the foregoing video-based human action recognition method can be performed by an action recognition composition architecture, for example, the composition architecture runs in the foregoing video processing device 300.

[0055] FIG. 4 shows a schematic diagram of a composition architecture of an action recognition scene according to the present application. The composition architecture 40 includes a video clip sequence 41, a human detection branch 42, an action detection model 43, a semantic segmentation model 44, and a fusion branch 45, and a video clip sequence action recognition result 46.

[0056] Each part of the composition architecture 40 of the action recognition scene is specifically described below.

[0057] The human detection branch 42 is configured to detect the human body in consecutive video frames of the input video clip sequence 41. The human detection branch 42 includes a human detection model configured to generate a bounding box of the human body, and an optical flow smoothing branch configured to correct the bounding box generated by the human detection model. Specifically, the human detection model first classifies the objects in the video frame to obtain a region where the human instance is located in the image, then determines the position of the region where the human instance is located in the video frame, and annotates the bounding box of the human instance at the corresponding position. The optical flow smoothing branch is configured to detect a plurality of image frames of the human bounding box for the human detection model, and optimize these bounding boxes by using the optical flow between the plurality of image frames, that is, adjust the positions of these bounding boxes, i.e., adjust the position of the detected human instance in the plurality of image frames.

[0058] In some other embodiments, the foregoing human detection branch is configured not only to detect the bounding box of the region where the human body is located in the image, but also to detect human-related objects in the image, for example, objects that can reflect human actions, such as the object "football" in contact with the foot of the human body can reflect the action of the human body as "playing football".

[0059] The action detection model 43 is configured to generate the action classification and probability distribution of the human instances in the video clip with reference to the interaction relationship between the target person's own motion of the video clip and other people, objects, and long-term memory information. The embodiments of the present application mainly take the test of a video to be detected in a crowded scene as an example, it is necessary to consider the interaction relationship between the person's own motion of the video clip and other people and long-term memory information. The action detection model 43 includes a video clip feature extraction branch 431 and a human action detection branch 432.

[0060] The video clip feature extraction branch 431 is a three-dimensional convolutional neural network for extracting representation information of the video clip. Specifically, the

SlowFast 8×8 ResNet-101 (SlowFast8x8-R101) network is used as the specific structure of the three-dimensional convolutional neural network. The network is first pre-trained on a Kinetics-700 data set, and the structure of the three-dimensional convolutional network is initialized using the pre-trained weight. The network performs a convolution operation by using fast and slow passages. The slow passage is used to extract spatial features in a plurality of image frames, and the fast passage is used to extract temporal motion features of the plurality of image frames, and then different dimensions of data of the plurality of image frames are processed by using different passages to obtain recognition information, and the behavior information of the video clip sequence 41 can be predicted according to the recognition information.

[0061] The human action detection branch 432 includes a human representation extraction module with a memory pool, a multi-interaction relationship modeling fusion network, three fully connected layers, and a sigmoid regression layer. Specifically, the human representation extraction module performs linear interpolation on each bounding box region on the behavior representations of a plurality of image frames by using RoIAlign, and a person's regional representation is obtained after pooling, and at the same time, the human representation extraction module obtains a memory representation through the memory pool, and the multi-interaction relationship modeling fusion network performs modeling fusion on the human body's regional representation and memory representation to obtain a robust behavior representation, and the predicted probability of each category is obtained through the fully connected layer and the sigmoid regression layer.

[0062] The semantic segmentation model 44 is configured to output a scene map and corresponding scene features. Specifically, the image depth features are first extracted through a convolutional neural network, and the extracted image features are divided into two parts: object features and scene features. Then, image segmentation is separately performed on the object features and the scene features, and then the two types of features subjected to segmentation are fused (superimposed) to obtain a final full-image scene parsing result. In the embodiments of the present application, since only the scene map and the scene features are required, the pixels and features of the object may not be output, and only the scene map and

corresponding scene features are output after the full-image scene parsing result is obtained.

[0063] The fusion branch 45 is configured to establish a coupling relationship between a plurality of action categories and a plurality of scenes for increasing the probability of the action category (output by the action detection model 43) coupled to the scene output by the semantic segmentation model 44. For example, the semantic segmentation model 44 outputs the stair scene, and at the same time, the action detection model 43 outputs an action of walking up and down stairs, a standing action, a bending over action, etc. and the corresponding probability distribution, where the action of walking up and down stairs establishes a coupling relationship of the stair scene. In this case, the probability of the action of walking up and down stairs is increased, and the probabilities of other actions are decreased after being normalized. Finally, the action of the human instance with the highest probability is output.

[0064] FIG. 5 shows a flowchart of a video-based human action recognition method according to some embodiments of the present application. The video-based human action recognition method provided in the present application is described in detail below with reference to FIGs. 1 to 5. As shown in FIG. 5, specifically, the method includes the following steps.

[0065] 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.

[0066] 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 action recognition method of the present application is used to perform human pose estimation on the video in the test set. The video-based human action recognition method of the present application is described herein by mainly taking action recognition of human instances in videos of crowded scenes as an example. For example, the video to be processed is preprocessed to generate the video clip sequence 41 shown in FIG. 4.

[0067] 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 video clip sequence 41. For example, video framing software or video screenshot software can be used to perform a framing operation on the video to be processed, specifically, the framing operation is performed on the video to be processed through
5 a videocapture function of an OpenCV library in the software.

[0068] For example, the duration of the video to be processed is 15 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 $60 \times 15 \times 60 = 54,000$ images, and it can be understood that 36,000 images are obtained after 10
10 minutes of video framing. "*" indicates multiplication.

[0069] In the embodiments of the present application, preprocessing is performed on the video to be processed, and the video clip sequence 41 is obtained after framing. Since the video to be processed is usually long or the video to be processed is a real-time video stream, it is impossible to obtain the action features of the entire video to be processed. As shown in Figure
15 6, The framed video to be processed can be divided into video clip sequences according to a time sequence with a preset number of m video frames. Specifically, the video clip sequence 41 can be obtained in the following manner: preprocessing the video to be processed; dividing the video frame of the video to be processed according to a preset number of m video frames after framing, to obtain a video clip containing m video frames; and forming the video clips obtained
20 through division into a video frame sequence according to the time sequence of the video frame. A middle frame of the video clip is extracted as the key frame.

[0070] In the embodiments of the present application, taking the video to be processed of the stair scene as an example, as shown in FIG. 6, the video is equally divided into n video clips, and the video clips form video clip sequences to be processed $\{P_1, P_2, \dots, P_t, \dots, P_n\}$, where P_t
25 is the t-th video clip, and the t-th video clip is a video clip that is adjacent to and located in front of the (t+1)-th video clip.

[0071] During specific implementation, the preset number can be used as a reference for adding action annotations every few frames in the training sample set for training the action

recognition model. For example, if the training sample set adds an action annotation to all individuals every 20 frames, then the preset number is 40 frames. The preset number can also be dynamically determined according to the length of the video to be processed. When the video to be processed is longer in length, the value of the preset number can be larger, and when the video to be processed is shorter in length, the value of the preset number can be smaller. For example, the preset number is positively correlated with the length of the video to be processed.

[0072] It can be understood that the video clip sequence 41 is separately input into the human detection branch 42 for human detection, to the semantic segmentation model 44 for scene extraction, and to the video clip feature extraction branch 431 for representation information extraction.

[0073] Step 504: The human detection model generates initial human bounding boxes of human instances in each multi-human image frame.

[0074] The human detection model can achieve three tasks. The first task is a classification task, which is to traverse the input consecutive video frames and classify the human bodies and non-human bodies of the video frames. The second task is a regression task, which is to determine the position of the human region in the video 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 video frame shown in FIG. 7, the human detection model 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 video frame.

[0075] Specifically, the human detection model generates the initial human bounding box of the human instance in the video frame (i.e., each multi-human image frame), where 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 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 video frame in the video clip sequence.

[0076] In the embodiments of the present application, the process of the human detection model performing human detection on the multi-human video frames in the input video clip sequence 41 includes: obtaining the position of each human body in the video frame, and predicting the bounding box of the human body for the region where each human body is located. Specifically, the video frame can be input into the pre-trained human detection model for human detection, the region of the human body in the video frame is obtained, and the human detection model predicts the bounding box of the human body for the region where each human body is located when human detection is performed on the video frame. The human detection model can be a pre-trained deep learning convolutional neural network. The human detection model 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 the video frame, so that the human detection model outputs the predicted bounding box of the human body more accurately.

[0077] 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 video clip sequence 41 after framing of the video to be processed, the video frames are sequentially input into the trained human detection model to obtain the initial human bounding box output by the human detection model.

[0078] Step 506: The optical flow smoothing branch 43 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.

[0079] 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 video frame. The optical flow smoothing corrects an initial human bounding box of the multi-human video frame based on position difference (i.e., displacement) data between the multi-human video frame and the previous video frame of the multi-human video frame, to obtain the detected human bounding box, where the position difference data includes a position difference of the human instance in the multi-human video frame between the previous video frame and the multi-human video frame. 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.

[0080] 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 and the initial human bounding box of the previous frame output by the human detection model, 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 video 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 video frames, thereby improving the accuracy of the finally detected human bounding box, that is, improving the accuracy of human action recognition.

[0081] 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 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 through optical flow smoothing is used for subsequent input into the action recognition model 43 for processing.

[0082] 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 video frame of the video clip through the following formula as an example, it is illustrated that the optical flow smoothing is performed on the bounding box of the human body detected by the human detection model in the embodiments of the present application. For example, the t -th video frame is the multi-human video frame in step 506.

[0083] FIG. 8 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 video frame. Specifically, as shown in the drawing, the coordinate set $\tilde{I}_i^k$ detected by the human bounding box I1 includes a point $w(x_1, y_1)$ at the upper right corner of the bounding box of the human body and a point $z(x_2, y_2)$ at the lower left corner of the bounding box of the human body. It can be understood that by obtaining the point $w(x_1, y_1)$ at the upper right corner and the point $z(x_2, y_2)$ at the lower left corner of the human bounding box I1, the region of the human instance I1 in the multi-human video frame can be obtained.

[0084] In a possible implementation, the embodiments of the present application calculate the detected coordinate set $\tilde{I}_i^k$ obtained through optical flow smoothing of the human bounding box I1 by using the following formula (1):

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

[0085] In the formula, in $\tilde{I}_i^t$, $\hat{I}_i^t$, and I_i^t , t represents the t -th video frame, and i represents the bounding box of the i -th human body in the t -th video frame.

[0086] α represents a weight coefficient, which is an empirical parameter; in most cases,

the value is 0.5.

[0087] $\hat{I}_i^t$ represents the coordinate set of the human bounding box I1 in the t-th video frame obtained through prediction according to the optical flow from the (t-1)-th frame to the t-th frame.

5 [0088] I_i^t represents the coordinate set of the initial human bounding box I1 in the t-th video frame output by the human detection model.

[0089] It can be understood that Formula (1) above can realize that the coordinate set $\hat{I}_i^t$ of the human bounding box I1 of the t-th video frame predicted according to the optical flow from the (t-1)-th frame to the t-th frame and the coordinate set I_i^t of the initial human bounding box I1 of the t-th frame output by the human detection model jointly determine the coordinate set $\tilde{I}_i^t$ for detecting the detected human bounding box I1 in the t-th video frame.

10

[0090] In addition, some embodiments of the present application can implement "the coordinate set $\hat{I}_i^t$ of the human bounding box I1 of the t-th video frame predicted according to the optical flow from the (t-1)-th frame to the t-th frame" by using Formula (2):

15
$$\hat{I}_i^t = I_i^{t-1} + F^{t-1 \rightarrow t} \quad (2)$$

[0091] In the formula, $F^{k-1 \rightarrow k}$: 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.

[0092] I_i^{t-1} represents the coordinate set of the initial human bounding box I1 in the (t-1)-th frame output by the human detection model.

20 [0093] Step 508: The video clip feature extraction branch 431 obtains behavior representation information of the video clip.

[0094] In the embodiments of the present application, the action detection model 43 includes a video clip feature extraction branch 431 (i.e., a three-dimensional convolutional neural network) and a human action detection branch 432. Specifically, the action detection model can be an asynchronous interaction aggregation (AIA) model, where the backbone network of the AIA model 43 is a pre-trained 3D convolutional neural network (3D CNN). The backbone network can be an inflated 3D conventional network (I3D) or a SlowFast 8x8 ResNet-101 (SlowFast8x8-R101) network. In the embodiments of the present application, the backbone

25

network of the AIA model adopts the SlowFast8x8-R101 network pre-trained on the Kinetics-700 data set. The SlowFast8x8-R101 network is a video recognition network based on the combination of fast and slow passages. The network performs 3D convolution operations immediately after capturing several frames with the fast and slow passages. It can be understood
5 that the slow passage (i.e., high resolution CNN) is used to analyze the spatial features in the video clip, while the fast passage (i.e., low resolution CNN) is used to analyze the motion features in the video. The behavior actions in the video are classified with reference to the features extracted from the two passages.

[0095] Specifically, the video to be processed is preprocessed, and framed to obtain video
10 clip sequences 41, and the video clip sequences 41 are sequentially input into the SlowFast8x8-R101 network. The video frames corresponding to the video clips are respectively input into the modules corresponding to the slow passage and the fast passage, and then different dimensions of data in the video frame are processed using different passages to obtain recognition information, and then the behavior representation information of the video clip
15 sequence 41 can be predicted according to the recognition information, and moreover, the human detection branch 42 generates the human bounding box, and the human action detection branch 432 obtains the behavior representation information in the human bounding box. The behavior representation information includes comprehensive information about human actions and timing in video frames, specifically referring to changes in human actions over time.

[0096] Specifically, FIG. 9 provides a schematic diagram of a fast and slow passage
20 connection mode based on the SlowFast8x8-R101 network. As shown in Figure 9 and Figure 6, for example, Divide the video to be processed according to the preset number m of video frames, the preset number m is 40, and a video clip in the video clip sequence includes 40 video frames. Taking a video clip P_t as an example, the video clip P_t samples 8 frames and 32 frames
25 to be respectively input into the slow passage and the fast passage, and then different dimensions of data in the video frame are processed using different passages to obtain recognition information for predicting the behavior representation information of the video clip sequence.

[0097] Step 510: The human behavior representation is extracted according to the behavior representation information of the video clip generated by the video clip feature extraction branch 431 and the bounding box of the human instance generated by the human detection branch 42, and the action category and probability distribution of the human instance are
5 predicted by a multi-interaction modeling fusion network, three connect layers and a sigmoid regression layer.

[0098] In the embodiments of the present application, the AIA model 43 includes the video clip feature extraction branch 431 (i.e., the three-dimensional convolutional neural network) and the human action detection branch 432. Specifically, the human detection branch 42
10 generates the bounding box of the region where the human body is located of the video frame and the video clip feature extraction branch 431 extracts the behavior representation information of the video clip as the input of the human action detection branch 432. The human action detection branch 432 obtains the person's regional representation and memory representation using the bounding box of the human instance after training, and multi-category
15 determination is further performed on this representation. The human action detection branch 432 includes a human representation extraction module 4321 with a memory pool, a multi-interaction relationship modeling fusion network 4322, three fully connected layers, and a sigmoid regression layer 4323.

[0099] FIG. 10 provides a schematic diagram of action recognition based on an AIA model.
20 The method for predicting the human action category by the AIA model is described below with reference to FIG. 10.

[00100] In the embodiments of the present application, the human action detection branch 432 includes a human representation extraction module 4321 with a memory pool, a multi-interaction relationship modeling fusion network, three fully connected layers, and a sigmoid
25 regression layer 4323. The representation extraction module 4321 performs linear interpolation on each bounding box region on the behavior representations by using RoIAlign, and a human body's regional representation is obtained after pooling, and at the same time, the representation extraction module obtains a memory representation through the memory pool, and the multi-

interaction relationship modeling fusion network performs modeling fusion on the regional representation and memory representation of the human body and objects to obtain a robust behavior representation, and the predicted probability of each category is obtained through the fully connected layers and the sigmoid regression layer.

5 **[0100]** In the embodiments of the present application, the video clip is input into the video clip feature extraction branch 431 to obtain the representation of the video clip combined with the bounding box of the region where the human body is located generated by the human detection branch 42, and the representation extraction module 4321 is used. According to the bounding box of the human body, further pooling is performed on the behavior representation
10 of the video clip by using the RoIAlign, to obtain the regional representation of the person. The video clips that generate the regional representation of the person are written into the memory pool in sequence, as shown in FIG. 10, the video clip sequence obtained from the regional representation of the person is $\{P'_1, P'_2, \dots, P'_t, \dots, P'_n\}$.

[0101] For example, action recognition is performed on a video clip P_t , and $2L+1$ video
15 clips are read from the memory pool, where L is an integer greater than or equal to 1. Specifically, the read $2L+1$ video clips include L video clips that are adjacent in front of and behind P'_t , and are spliced into a memory representation sequence $\{P'_{t-L}, \dots, P'_t, \dots, P'_{t+L}\}$ of the video clip P_t ; the person's regional representation and the person's memory representation are input into the multi-interaction relationship modeling fusion network, and modules in the
20 multi-interaction relationship modeling fusion network are respectively configured to model different interaction relationships to obtain a more robust behavior representation of the video clip P'_t .

[0102] Specifically, FIG. 11 is a schematic structural diagram of a multi-interaction relationship modeling fusion network 4322. As shown in FIG. 11, the multi-interaction
25 relationship modeling fusion network 4322 includes: two human-human interaction modeling modules configured to receive the regional representation of the person, and two human-memory modeling interaction modules configured to respectively receive the regional representation and memory representation of the person, where the first human-human

interaction modeling module, the first human-memory modeling interaction module, the second human-human interaction modeling module, and the second human-memory modeling interaction module are sequentially connected to each other and transmit the regional representations of the people that are sequentially enhanced. Each interaction modeling module models one of the human-human interaction and the human-memory interaction, and is transmitted to the next module after being fused with the regional representation of the person. The finally output regional representation of the person integrates the human-human interaction and human-memory interaction, which is the finally output robust behavior representation. The human-human interaction refers to the interaction between different actors in the video clip P_t . The human-memory interaction refers to the interaction between actors in the memory representation sequence $\{P'_{t-L}, \dots P'_t, \dots P'_{t+L}\}$ of the video clip P_t in the video clip P_t .

[0103] Specifically, the modeling refers to:

$$h(Q, K) = QW_Q + \left[\text{Softmax} \left(\frac{QW_Q W_{K1}^T K^T}{\sqrt{d}} \right) KW_{K2} \right] W_0$$

where Q and K represent two representations of the input, respectively, W_Q , W_{K1} , W_{K2} , and W_0 represent the weights of the fully connected layers, and d represents the dimension of KW_{K1} . According to the different input representations K , the module processes different interaction relationships. For example, the value of K includes person's regional representations and memory representations. The corresponding modeling modules sequentially process human-human interaction and human-memory interaction accordingly, and output corresponding representations fused with this type of interaction information. The output of the previous modeling module is further used as the input of the next Q when the four modules are connected in series, and finally a variety of different interaction relationships are fused.

[0104] In the embodiments of the present application, as shown in FIG. 10, three fully connected layers and a sigmoid regression layer 4323 obtain the predicted probabilities of each category. The three fully connected layers include two hidden layers and one output layer. The sigmoid regression layer includes a sigmoid function and a cross-entropy loss function. An

output vector of the output layer can obtain the predicted probability of each category by using the sigmoid layer. The cross-entropy loss function is used to train the entire network. The training here refers to: taking the samples in the sample set, the corresponding object bounding boxes, and the regional representations of the person close to the video clips in the memory pool disposed in the representation extraction module 4321 as the input of the behavior detection network, adjusting the network parameters by using the cross-entropy loss function combined with the back propagation (BP) algorithm, and updating the regional representation of the person in the video clip to the memory pool.

[0105] In the embodiments of the present application, the action detection model 43 includes the video clip feature extraction branch 431 (i.e., the three-dimensional convolutional neural network) and the human action detection branch 432. The action detection model 43 is a deep learning convolutional neural network obtained through training based on a large-scale training sample set having action annotations of human instances.

[0106] Step 512: The semantic segmentation model 44 determines the image scene of each multi-human image frame.

[0107] In the embodiments of the present application, the semantic segmentation model 44 extracts scenes in video clips, where the semantic segmentation model 44 is a pre-trained convolutional neural network. Specifically, the depth features of the video frame are first extracted through the convolutional neural network, and the extracted image features are divided into two parts: object features and scene features. Then, image segmentation is separately performed on the object features and the scene features, and then the two types of features subjected to segmentation are fused (superimposed) to obtain a final full-image scene parsing result. The scene map and corresponding scene features are output after the full-image scene parsing result is obtained. FIG. 12a is a visualized schematic diagram of a stair scene output by the semantic segmentation model 44, The left image of FIG. 12a is a scene diagram of the stair scene extraction after the semantic segmentation model 44 removes the human instance in the video frame, and the right image of FIG. 12a is the stair position pixel point diagram of the stair scene after removing the human instance in the video frame. The stair

position pixel point map represents the stair position information of the stair scene. Figure. Figure 12b is a schematic diagram of the visualization of the passage scene output by the semantic segmentation model 44. The left image of FIG. 12b is a scene diagram of the passage scene extraction after the semantic segmentation model 44 removes the human instance in the video frame, and the right image of FIG. 12b is the stair position pixel point diagram of the passage scene after removing the human instance in the video frame. The stair position pixel point map represents the stair position information of the passage scene. In the embodiments of the present application, the semantic segmentation model 44 is a deep convolutional neural network trained by a sample set labeled with a semantic region tag. For example, the semantic segmentation model is pre-trained using an ADE20K data set. ADE20K is an image segmentation data set in a complex scene, including 20,000 images as a training set, 2,000 images as a verification set, and 3,000 images as a test set.

[0108] In the embodiments of the present application, common backbone deep convolutional neural networks for semantic segmentation are, for example, AlexNet, VGG-16, GoogleNet, ResNet, and PSPNet, etc. PSPNet is taken as an example to illustrate the scene where the semantic segmentation model extracts the video frame. Specifically, a pyramid analysis network (PSPNet) is used as the semantic segmentation model 44 for implementing the convolutional neural network. The input is a video frame of the video clip sequence 41, and the output result is the scene category of the video frame. In order to achieve this process, the input video frame is first processed into a feature map by using PsNet, and then a pyramid pooling operation is performed on the generated feature map, so that feature maps of different resolutions can be obtained, and then a convolution operation is performed on each pooled feature map; with reference to up-sampling feature maps, the results are stacked to form the final feature map, and finally a score map of the object or scene category is obtained through convolution processing. The Softmax activation function is used during the category output mapping to generate a probability distribution (the total score is 1). It can be understood that taking a scene category score map as an example refers to a probability distribution map corresponding to a scene category that may be output by a video frame. For example, the

probability of the passage scene is 0.5, the stair scene is 0.4, and the subway entrance/exit scene is 0.1. Then, according to the semantic fusion method, the semantic tag with the highest probability is selected for each pixel. These probabilities are referred to as the semantic confidence level of the associated semantic category tag here. It can be understood that the semantic segmentation model 44 outputs the scene category with the highest confidence level and the object category with the highest confidence level. The scene category is used as the input of the fusion branch 45.

[0109] Step 514: The fusion branch 45 recognizes the actions of human instances in the plurality of multi-human image frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames.

[0110] In the embodiments of the present application, the action recognition of the human instance is performed on the video clip sequence 41. First, the human detection branch 42 performs human detection on the content of the video frame (i.e., the plurality of multi-human image frames), and generates the bounding box of the human body; the action detection model 43 predicts the action category and probability of the human instances with reference to the bounding box of the human body, and then post-processing is performed on the predicted action category and probability of the human instances with reference to the environment features of the human instances in the video clip, that is, weighted processing is performed again on the action category and probability of the human instances predicted by the action detection model 43, to increase the weight of the action category related to environment features. The environment features include the scene category (the stair scene, the restaurant scene, etc.) in the video frame, the interaction between people (i.e., fighting, running together, gathering, etc.), the interaction between people and objects (a person pushing a cart, taking a cup to drink water, etc.) etc.

[0111] Specifically, the bounding box IoU of two human instances in the video frame may be higher when the two human instances move together or fight or gather together. Weighted processing is further performed on the action category and probability distribution output by the action detection model 43 according to the IoU. For example, during action recognition of

the video clip sequence, the action category and probability distribution of the human instance predicted by the action detection model 43 are: The probability of running together is 0.6, the probability of standing is 0.2, the probability of fighting is 0.2, where the action categories, i.e., running together and fighting require interaction between two or more people; the fusion branch 45 further determines whether there is a human instance in the bounding box with a higher degree of overlap where the human instance is located, and if not, Then the fusion branch 45 re-weights the action category and probability of the human instance, that is, the probability of running together is reduced by 0.5, and the probability of fighting is reduced by 0.5. Then the probability distribution of the action category of the human instance becomes: The probability of running together is 0.1, the probability of standing is 0.2, and the probability of fighting is - 0.3; Then perform the softmax operation on the probabilities of all action categories of the weighted human instance to ensure that the overall probability is still 1. Then the action category and probability distribution of the redistributed weights are respectively the probability of running together 0.36, the probability of standing 0.40, and the probability of fighting 0.24. Specifically, the fusion branch 45 performs post-processing on the predicted action category and probability of the human instances with reference to the scene of the human instances in the video clip, that is, weighted processing is further performed on the action category and probability of the human instance predicted by the action detection model 43, to increase the weight of the action category related to the environment features. When there is a stair scene, the action of the person on the stairs is strongly related to the stair scene. Therefore, the fusion branch 45 establishes a coupling relationship between all action categories and scenes that may be output by the action detection model 43. One image scene corresponds to one or more categories of actions, or a plurality of image scenes correspond to one or more categories of actions.

[0112] For example, FIG. 13 shows a schematic diagram of the coupling relationship established between the scene category and the action category. As shown in FIG. 13, the action categories that may be output by the action classification in the action detection model 43 include: walking up and down the stairs, running, backpack stealing, fighting, etc. The scene

categories that may be output by the semantic segmentation model 44 include crossroads, restaurants, stairs, waiting rooms, etc. The arrow indicates the coupling relationship between the action category and the scene category. The action category, i.e., walking up and down the stairs establishes a coupling relationship with the stair scene and the passage scene. The action category, i.e., running establishes a coupling relationship with the crossroad scene and the waiting room scene. The action category, i.e., backpack stealing establishes a coupling relationship with the restaurant scene, the waiting room scene, and the bus scene. The action category, i.e., fighting establishes a coupling relationship with the restaurant scene, the bus scene, and the passage scene. The action category, i.e., standing establishes a coupling relationship with the stair scene and the passage scene. The action category, i.e., falling over establishes a coupling relationship with the stair scene, the crossroads scene, and the passage scene, and so on. The coupling relationship here is used to perform weighted processing on the action category and probability of the human instance in the bounding box of the video clip output by the action detection model 43 to increase the probability of the action category coupled with the scene output by the semantic segmentation model 44. It can be understood that this is only an example, the action category output by the action detection model 43 and the scene category output by the semantic segmentation model 44 are not limited thereto, and the coupling relationship between the action category and the scene category is not limited thereto, either.

[0113] Specifically, taking action recognition of a video of the stair scene as an example, for example, when the scene output by the semantic segmentation model 44 is the stair scene, the action category and probability distribution of human instance output by the action detection model 43 are that the probability of walking up and down the stairs is 0.4, the probability of backpack stealing is 0.4, and the probability of fighting is 0.2, respectively. It can be understood that the action categories, i.e., walking up or down the stairs establish a coupling relationship with the stair scene, then the weights are reallocated to the action category and probability. For example, the probability of the action category, i.e., walking up and down the stairs is increased by 0.5, Then the probability distribution of the action category of the human instance becomes:

the probability of walking up and down stairs is 0.9, the probability of backpack stealing is 0.4, and the probability of fighting is 0.2; and then a softmax operation is performed on the probabilities of all actions of human instance to ensure that the sum of the overall probabilities is still 1. Then the action category and probability distribution of the redistributed weights are
5 the probability of walking up and down stairs 0.46, the probability of backpack stealing is 0.3, the probability of fighting is 0.24, and walking up and down stairs with the highest probability are taken as the final output action.

[0114] In the action recognition solution of the present application, the human body in the video frame in the video clip sequence 41 is detected by the human detection model to generate
10 a bounding box of the human body. Specifically, the function and training process of the model are as follows.

[0115] (1) Training purpose of the human detection model

[0116] 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, and the
15 performance of the human detection model 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
20 embodiments of the present application, the larger the AP and the smaller the MMR is, the better the performance of the human detection model is.

[0117] (2) Selection of a model framework of the human detection model

[0118] 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 may be a one-stage
25 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.

[0119] 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.

[0120] The framework selection results in the human detection model are evaluated by using table 1 below.

[0121] Table 1 shows the experimental results of the human detection model 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 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. 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 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.

[0122] It can be seen from Table 1 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 1		
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

[0123] (2) Selection of a training sample data set for the human detection model

[0124] 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. The human training sample set for training the human detection model 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 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. 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.

5 **[0125]** The selection and verification results of the training sample set for training the human detection model are evaluated by using table 2 below.

10 **[0126]** Table 2 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 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.

15 **[0127]** As can be seen from Table 2, 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
20 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 CrownHuman data set and the self-collected data set, the human detection performance has the AP increased by 0.33 and the MMR
25 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 in the embodiments of the present application

includes the HIE data set, the CrowndHuman data set, and the self-collected data set.

Table 2		
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

[0128] (3) Training details of the human detection model

[0129] 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.

[0130] 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 3		
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

[0131] Table 3 shows the verification result that the human detection model is the baseline model, the baseline model is trained on the training sample set, and the human detection test is performed on the HIE validation set. The baseline model is a human detection model 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.

[0132] It can be seen from Table 3 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.

[0133] (4) KD regularization of the human detection modelIn the embodiments of the present application, in the process of performing human detection on each image frame by the

human detection model, the output layer of the human detection model 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 selects the bounding box with the highest category probability of the human body. In order to reduce the overfitting of the human detection model 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.

[0134] 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_{τ} represents the probability distribution of the category output by the detection model after temperature t soften, p^d_{τ} represents the artificially designed probability distribution, and the distribution formula is as follows:

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

[0135] Finally, based on the analysis above, the HIE data set, the crowdhuman data set and the self-collected data set are adopted to train two human detection models, 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 in the embodiments of the present application.

[0136] In the action recognition solution of the present application, action recognition is performed on the human body in the video clip by the action detection model 43, to generate the action category of the human body. Specifically, taking the action detection model 43 being an AIA model as an example, the function and training process of the model are as follows.

[0137] (1) Selection of the action training sample data set of the human action recognition model

[0138] In the embodiments of the present application, the training sample set for training the AIA model includes an AVA data set and an HIE data set.

[0139] The AVA data set contains 430 15-minute annotated Youtube videos. The 15-minute videos are uniformly divided into 300 non-overlapping 3-second clips. The sampling follows the time sequence of ensuring the action sequence, and the human and objects in the middle frame of each 3-second video clip are manually annotated with bounding boxes and actions. Specifically, the AVA data set has 80 atomic visual actions(80 action categories), and the annotation of each action includes the descriptive text, which is also included in the bounding

box in the video frame. These actions can be categorized into three groups: pose/movement actions, human-object interaction, and human-human interaction. For example, all actions such as walking, kicking (something), shaking hands, and running have spatio-temporal positioning, resulting in 57.6k video clips, 96k annotated human actions, and 232k action tags.

5 **[0140]** The HIE training set contains 33559 human action instances from 14 action categories. Action annotations are added to all individuals every 20 frames, and the frames containing action annotations are called key frames. However, the distribution of all action categories in the HIE data set is unbalanced. Running alone, running together, sitting down and talking, fighting and falling each have less than 400 human instances, and 11.7k actions are
10 annotated as walking alone. Moreover, For a video frame, the area where all human instances are located, on average, only accounts for less than 2% of the frame area of the entire video frame. Therefore, performing spatio-temporal action localization on this data set is a challenging task. In order to alleviate the problem of insufficient data, some additional data has been added. The additional data contains 3k human instances of 2k video frames, and 2.9k
15 action annotations are added to the human instances.

[0141] (2) Training details of the human detection model

[0142] (a) Data set splitting

[0143] It is attempted to split the data set into training and verification sets through videos, similar to the detection training. However, there are few positive instances of several action tags
20 in the training and/or verification set, which makes it difficult to evaluate the effect with this splitting strategy. Therefore, several rounds of initial experiments are conducted by using the train verification splitting above to verify the correctness of our implementation. Hereafter, it is switched to the training of a complete training set that is not verified, and then the given official test set is tested.

25 **[0144]** (b) Training details

[0145] For each trim clip around the key frame, the slow branch and the fast branch in a feature extractor are sampled in 32 frames and 8 frames respectively in the clip. For the spatial size, in the training process, the short side of the framework is adjusted to the length randomly

sampled from (320, 352, 384, 416, 448), and the long side is filled to 2.5 times the short side. 10% bounding box jitter and random horizontal flip are applied to improve the generalization ability. A dropout rate of 0.2 is used before classification prediction, and a binary cross-entropy loss of an S-type activation log is adopted as supervision in the training process. A total of 160,000 iterations of training are performed using an SGD optimizer, in which BatchNorm statistical information is frozen. Each model is trained on eight GPUs, each having two clips, which brings the total number of batches to 16.

[0146] (c) Weighted sampling

[0147] Due to the lack of training samples for some courses, it is observed that the score distribution of rare courses after training is close to zero. In order to reduce the problem of imbalance between action categories, weighted sampling is performed on the clips in the training process, and higher weights are allocated to clips that contain rare actions (such as sitting and running), and lower weights are allocated to clips that contains frequent actions (such as walking and standing). Upon application of a weighted sampling strategy, in the training process, the ratio of the number of occurrences of the maximum sampling action to the number of occurrences of the minimum sampling action is reduced from about 120 to about 10.

[0148] FIG. 15 shows a structural block diagram of a video-based human action recognition apparatus according to some embodiments of the present application. As shown in FIG. 15, the video-based human action recognition apparatus 1500 includes:

an acquisition module (1502), 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 first generation module (1504), configured to generate detected human bounding boxes of human instances in each multi-human image frame; a determining module (1506), configured to determine an image scene of each multi-human image frame, one image scene corresponding to one or more categories of actions; and a recognition module (1508), configured to recognize the actions of human instances in the plurality of multi-human image frames based on the detected human bounding boxes and the image scenes of the plurality of multi-human image frames. It can be understood that the video-based human action recognition apparatus 1500

shown in FIG. 15 corresponds to the video-based human action recognition method provided in the present application. The above technical details in the specific description of the video-based human action recognition method provided in the present application are still applicable to the video-based human action recognition apparatus 1500 shown in FIG. 15. Reference may
5 be made to the specific description above, and details are not described herein again.

[0149] FIG. 16 shows a block diagram of an electronic device 1600 according to one embodiment of the present application. The electronic device may be the foregoing video processing device 300. FIG. 16 schematically illustrates an exemplary electronic device 1600 according to a plurality of embodiments. In one embodiment, the electronic device 1600 may
10 include one or more processors 1604, a system control logic 1608 connected to at least one of the processors 1604, a system memory 1612 connected to the system control logic 1608, a non-volatile memory (NVM) 1616 connected to the system control logic 1608, and a network interface 1620 connected to the system control logic 1608.

[0150] In some embodiments, the processor 1604 may include one or more single-core or
15 multi-core processors. In some embodiments, the processor 1604 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 1600 adopts an evolved Node B (eNB) or radio access network (RAN) controller, the processor 1604 may be configured to execute conforming embodiments, e.g.,
20 one or more of the plurality of embodiments shown in FIG. 3.

[0151] In some embodiments, the system control logic 1608 may include any suitable interface controller to provide any suitable interface to at least one of the processors 1604 and/or any suitable device or component in communication with the system control logic 1608.

[0152] In some embodiments, the system control logic 1608 may include one or more
25 memory controllers to provide an interface connected to the system memory 1612. The system memory 1612 can be configured to load and store data and/or instructions. In some embodiments, the memory 1612 of the system 1600 may include any suitable volatile memory, such as a suitable dynamic random access memory (DRAM).

[0153] The NVM/memory 1616 may include one or more tangible, non-transient computer readable media for storing data and/or instructions. In some embodiments, the NVM/memory 1616 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.

[0154] The NVM/memory 1616 may include a part of the storage resources on an apparatus where the electronic device 1600 is installed, or it may be accessed by the device, but not necessarily a part of the device. For example, the NVM/memory 1616 can be accessed through the network via the network interface 1620.

[0155] In particular, the system memory 1612 and the NVM/memory 1616 may respectively include: a temporary copy and a permanent copy of the instruction 1624. The instruction 1624 may include an instruction that, when executed by at least one of the processors 1604, causes the electronic device 1600 to implement the method shown in FIG. 5. In some embodiments, the instruction 1624, hardware, firmware, and/or software components thereof may additionally/alternatively be placed in the system control logic 1608, the network interface 1620, and/or the processor 1604.

[0156] The network interface 1620 may include a transceiver, configured to provide a radio interface for the electronic device 1600, 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 1620 may be integrated in other components of the electronic device 1600. For example, the network interface 1620 may be integrated in at least one of the processor 1604, the system memory 1612, the NVM/memory 1616, or a firmware device (not shown) with instructions. The electronic device 1600 implements the method shown in FIG. 5 when at least one of the processors 1604 executes the instruction.

[0157] The network interface 1620 may further include any suitable hardware and/or firmware to provide a multi-input multi-output radio interface. For example, the network interface 1620 may be a network adapter, a wireless network adapter, a telephone modem and/or a wireless modem.

[0158] In one embodiment, at least one of the processors 1604 may be packaged together with the logic of one or more controllers for the system control logic 1608 to form a system in package (SiP). In one embodiment, at least one of the processors 1604 may be integrated on the same die with the logic of one or more controllers for the system control logic 1608 to form a system on chip (SoC).

[0159] The electronic device 1600 may further include an input/output (I/O) device 1632. The I/O device 1632 may include a user interface which enables a user to interact with the electronic device 1600. The design of a peripheral component interface enables a peripheral component to interact with the electronic device 1600. In some embodiments, the electronic device 1600 further includes a sensor, configured to determine at least one of environment conditions and position information related to the electronic device 1600.

[0160] 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.

[0161] 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.

[0162] 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 1620 or interact with the network interface 1620 to communicate with components of the positioning network (e.g., a global positioning system (GPS) satellite).

[0163] According to the embodiments of the present application, FIG. 17 shows a block diagram of a SoC 1700. In FIG. 17, similar components have the same reference numerals. In addition, the dashed-line box is an optional feature of more advanced SoCs. In FIG. 17, the SoC 1700 includes: an interconnection unit 1750, coupled to an application processor 1710; a system agent unit 1770; a bus controller unit 1780; an integrated memory controller unit 1740; a group of or one or more coprocessors 1720 which may include an integrated graphics logic, an image

processor, an audio processor, and a video processor; a static random access memory (SRAM) unit 1730; and a direct memory access (DMA) unit 1760. In one embodiment, the coprocessor 1720 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.

5 **[0164]** 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),
10 at least one input device and at least one output device.

[0165] 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
15 signal processor (DSP), a microcontroller, an application-specific integrated circuit (ASIC), or a microprocessor.

[0166] 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
20 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.

[0167] In some cases, the disclosed embodiments may be implemented in hardware, firmware, software, or any combination thereof. The disclosed embodiments can also be
25 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.

[0168] 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.

[0169] 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.

[0170] 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
5 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.

10 **[0171]** 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 action recognition method, comprising:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

t are positive integers.

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

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

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

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