TARGET TRACKING METHOD AND APPARATUS, ELECTRONIC DEVICE, AND STORAGE MEDIUM

Technical Field

The present disclosure relates to computer vision technologies, and in particular, to a target tracking method and apparatus, an electronic device, and a storage medium.

Background

In recent years, with continuous development of an artificial intelligence technology, the artificial intelligence technology achieves good effects in aspects such as computer vision and speech recognition. When target recognition and tracking are performed in some relatively special scenes (such as a tabletop game scene), it would be easy to cause misjudgment during recognition on a single frame of picture, while too large deep learning model during recognition on multiple frames of videos would cause high system time consumption.

Summary

The present disclosure provides solutions for target tracking.

According to one aspect of the present disclosure, provided is a target tracking method. The method includes: detecting a to-be-processed image to obtain a hand detection result; in response to the hand detection result including a bounding box of hand, determining a hand with a hand pose in the bounding box of hand conforming to a hand pose in a target gesture as a target hand; and tracking the target hand in a video stream according to the target hand in the to-be-processed image, where images in the video stream and the to-be-processed image are obtained by capturing a same target area, and the images in the video stream are obtained by capturing subsequent to capture of the to-be-processed image.

According to any one of implementations provided by the present disclosure, the in response to the hand detection result including a bounding box of hand, determining a hand with a hand pose in the bounding box of hand conforming to a hand pose in a target gesture as a target hand includes: in response to the hand detection result including a bounding box of hand, determining a hand in the bounding box of hand, with the hand pose conforming to the hand pose in the target gesture and in a pre-defined area in the to-be-processed image, as the target hand.

According to any one of implementations provided by the present disclosure, determining that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture includes: in response to the hand detection result including the bounding box of hand, cropping an image located in the bounding box of hand in the to-be-processed image; and determining, according to

the cropped image, that a hand pose in the cropped image conforms to the hand pose in the target gesture.

According to any one of implementations provided by the present disclosure, the hand pose in the target gesture is that two hands both face up; determining that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture includes: inputting the image in the bounding box of hand to a hand classification neural network to obtain a hand classification result, the hand classification result including a hand facing up or the hand with other poses; and if classification results in two bounding boxes of hand are that two hands both face up, determining that the hand poses in the bounding boxes of hand conform to the hand pose in the target gesture.

According to any one of implementations provided by the present disclosure, the hand classification neural network is obtained by training with hand image samples, a hand image sample including an image sample with the hand facing up, and an annotation result of the image sample with the hand facing up being the palm of the hand.

According to any one of implementations provided by the present disclosure, determining that the hand in the bounding box of hand is located in the pre-defined area in the to-be-processed image includes: determining that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image.

According to any one of implementations provided by the present disclosure, the pre-defined area includes two set sub-areas corresponding to two sub-areas in the target area respectively in the to-be-processed image; the determining that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image includes: determining that the center points of the two bounding boxes of hand are located in the two set sub-areas respectively.

According to any one of implementations provided by the present disclosure, the tracking the target hand in a video stream according to the target hand in the to-be-processed image includes: for an image frame in the video stream, determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame, where the first matrix of the image frame includes position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

According to any one of implementations provided by the present disclosure, the determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame includes: in response to a hand detection result of the image frame not including the bounding box of hand of the target hand, determining

the first matrix of the previous image frame as the first matrix of the image frame, and correcting the second matrix of the previous image frame with a correction matrix to obtain the second matrix of the image frame, where the correction matrix represents an error of each element in the second matrix of the previous image frame.

According to any one of implementations provided by the present disclosure, the determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame includes: in response to the hand detection result of the image frame including the bounding box of hand of the target hand, predicting, according to a time interval between the image frame and the previous image frame, the first matrix and the second matrix of the image frame according to the first matrix of the previous image frame; obtaining a third matrix and a fourth matrix of the image frame according to the position information of the bounding box of hand of the target hand, where the third matrix includes the position information of the center of the bounding box of hand of the target hand in the image frame, and the fourth matrix represents an error of each element in the third matrix; and multiplying Gaussian distributions corresponding to the second matrix by Gaussian distributions corresponding to the fourth matrix to obtain new Gaussian distributions, updating the first matrix according to an average value of the obtained new Gaussian distributions, and updating the second matrix according to covariance of the obtained new Gaussian distributions.

According to any one of implementations provided by the present disclosure, the to-be-processed image is an image of the target area captured at an overhead view.

According to one aspect of the present disclosure, provided is a target tracking apparatus. The apparatus includes: a detecting unit, configured to detect a to-be-processed image to obtain a hand detection result; a determining unit, configured to: in response to the hand detection result including a bounding box of hand, determine a hand with a hand pose in the bounding box of hand conforming to a hand pose in a target gesture as a target hand; and a tracking unit, configured to track the target hand in a video stream according to the target hand in the to-be-processed image, where images in the video stream and the to-be-processed image are obtained by capturing a same target area, and the images in the video stream are obtained by capturing subsequent to capture of the to-be-processed image.

According to any one of implementations provided by the present disclosure, the determining unit is specifically configured to: in response to the hand detection result including the bounding box of hand, determine a hand in the bounding box of hand, with the hand pose conforming to the hand pose in the target gesture and in a pre-defined area in the to-be-processed image, as the target hand.

According to any one of implementations provided by the present disclosure, the determining unit is configured to: in response to the hand detection result including the bounding box of hand, crop an image located in the bounding box of hand in the to-be-processed image; and determine, according to the cropped image, that a hand pose in the cropped image conforms to the hand pose in the target gesture.

According to any one of implementations provided by the present disclosure, the hand pose in the target gesture is that two hands both face up; the determining unit is configured to: input the image in the bounding box of hand to a hand classification neural network to obtain a hand classification result, the hand classification result including a hand facing up or the hand with other poses; and if classification results in two bounding boxes of hand are that two hands both face up, determine that the hand poses in the bounding boxes of hand conform to the hand pose in the target gesture.

According to any one of implementations provided by the present disclosure, the hand classification neural network is obtained by training with hand image samples, a hand image sample including an image sample with the hand facing up, and an annotation result of the image sample with the hand facing up being the palm of the hand.

According to any one of implementations provided by the present disclosure, the determining unit configured to determine that the hand in the bounding box of hand is located in the pre-defined area in the to-be-processed image, is configured to: determine that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image.

According to any one of implementations provided by the present disclosure, the pre-defined area includes two set sub-areas corresponding to two sub-areas in the target area respectively in the to-be-processed image; the determining unit configured to determine that the center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image, is configured to: determine that the center points of the two bounding boxes of hand are located in the two set sub-areas respectively.

According to any one of implementations provided by the present disclosure, the tracking unit is configured to: for an image frame in the video stream, determine a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame, where the first matrix of the image frame includes position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

According to any one of implementations provided by the present disclosure, the tracking unit configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is configured to: in response to a hand detection result of the image frame not including the bounding box of hand of the target hand, determine the first matrix of the previous image frame as the first matrix of the image frame, and correct the second matrix of the previous image frame with a correction matrix to obtain the second matrix of the image frame, where the correction matrix represents an error of each element in the second matrix of the previous image frame.

According to any one of implementations provided by the present disclosure, the tracking unit configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is configured to: in response to the hand detection result of the image frame including the bounding box of hand of the target hand, predict, according to a time interval between the image frame and the previous image frame, the first matrix and the second matrix of the image frame according to the first matrix of the previous image frame; obtain a third matrix and a fourth matrix of the image frame according to the position information of the bounding box of hand of the target hand, where the third matrix includes the position information of the center of the bounding box of hand of the target hand in the image frame, and the fourth matrix represents an error of each element in the third matrix; and multiply Gaussian distributions corresponding to the second matrix by Gaussian distributions corresponding to the fourth matrix to obtain new Gaussian distributions, update the first matrix according to an average value of the obtained new Gaussian distributions, and update the second matrix according to covariance of the obtained new Gaussian distributions.

According to any one of implementations provided by the present disclosure, the to-be-processed image is an image of the target area captured at an overhead view.

According to one aspect of the present disclosure, provided is an electronic device, including a memory and a processor, where the memory is configured to store computer instruction runnable on the processor, and the processor is configured to implement the target tracking method according to any one of implementations of the present disclosure when executing the computer instructions.

According to one aspect of the present disclosure, provided is a computer-readable storage medium, having a computer program stored thereon, where when the program is executed by a processor, the target tracking method according to any one of the implementations of the present disclosure is implemented.

In the target recognition method and apparatus, electronic device, and storage medium according to one or more embodiments of the present disclosure, by detecting a hand pose in a bounding box of hand in a hand detection result to determine a target hand that conforms to a hand pose in a target gesture, i.e., recognizing a target hand to be tracked from a to-be-processed image, and tracking the target hand in a synchronously captured video stream, the target hand can be recognized quickly and effectively according to the target gesture, thereby improving tracking efficiency.

Brief Description of the Drawings

To describe the technical solutions in one or more embodiments of the description or the prior art more clearly, the accompanying drawings required for describing the embodiments or the prior art are briefly introduced below. Apparently, the accompanying drawings in the following description show merely some embodiments described in the one or more embodiments of the description, and a person of ordinary skill in the art can still derive other accompanying drawings from these accompanying drawings without involving an inventive effort.

FIG. 1 is a flowchart of a target tracking method provided by at least one embodiment of the present disclosure;

FIG. 2 is a schematic diagram of an application scene of a target tracking method provided by at least one embodiment of the present disclosure;

FIGS. 3A-3F are schematic diagrams of a hand classification result provided by at least one embodiment of the present disclosure;

FIG. 4 is a schematic diagram of an application scene of a target tracking method provided by at least one embodiment of the present disclosure;

FIG. 5 is a schematic structural diagram of a target tracking apparatus provided by at least one embodiment of the present disclosure; and

FIG. 6 is a schematic structural diagram of an electronic device according to at least one embodiment of the present disclosure.

Detailed Description

To make a person skilled in the art better understand the technical solutions in one or more embodiments of the description, the technical solutions in the one more embodiments of the description are clearly and fully described below with reference to the accompanying drawings in the one or more embodiments of the description. Apparently, the described embodiments are merely some of the embodiments of the description, but not all the embodiments. Based on the

one or more embodiments of the description, all other embodiments obtained by a person of ordinary skill in the art without involving an inventive effort shall fall within the scope of protection of the present disclosure.

At least one embodiment of the present disclosure provides a target tracking method. The target tracking method may be executed by an electronic device such as a terminal device or a server. The terminal device may be a User Equipment (UE), a mobile device, a user terminal, a terminal, a cellular phone, a cordless phone, a Personal Digital Assistant (PDA), a handheld device, a computing device, a vehicle-mounted device, a wearable device, or the like. The method may be implemented by a processor by invoking computer-readable instructions stored in a memory.

FIG. 1 shows a flow of the target tracking method. As shown in FIG. 1, the method may include steps 101-103.

In step 101, a to-be-processed image is detected to obtain a hand detection result.

In the embodiments of the present disclosure, the to-be-processed image may be an image of a tabletop area captured by an image capture device (for example, a camera). The image may include a complete tabletop image or may include a part of an image of a tabletop.

The image captured by the camera may be a frame in a video stream or may be an image acquired in real time, and no limitation is made thereto in the present disclosure.

In the embodiments of the present disclosure, a hand in the image may be detected via a deep learning network, for example, faster RCNN network, so as to obtain a hand detection result. A person skilled in the art should understand that the faster RCNN network is only an example, and other networks may also be used. No limitation is made thereto in the present disclosure.

In some embodiments, the to-be-processed image may be captured at an overhead view, i.e., the to-be-processed image is captured by a birdview camera. For an application scene, i.e., the tabletop, the top view is a good photography view, and the view can generally avoid blocking a detection target by other objects. However, a person skilled in the art should understand that the to-be-processed image captured by the camera at other views may also be applied to the target tracking method provided in the embodiments of the present disclosure.

In step 102, in response to the hand detection result including a bounding box of hand, it is determined that a hand with a hand pose in the bounding box of hand conforming to a hand pose in a target gesture as a target hand.

In the embodiments of the present disclosure, the image the bounding box of hand is detected, and a detected feature is compared with a feature of the hand pose in the target gesture to

determine whether the hand pose in the bounding box of hand conforms to the hand pose in the target gesture; or hand poses in the bounding box of hand are classified, and it is determined whether the hand poses conform to the hand pose in the target gesture by means of a classification result. The specific determining method would be described in detail below.

In response to the hand detection result including a bounding box of hand, for one or more bounding boxes of hand in the hand detection result, the bounding box of hand, where the hand pose conforms to the hand pose in the target gesture, may be determined as a target detection bounding box, and the hand in the target detection bounding box is determined as target hand.

In step 103, the target hand is tracked in a video stream according to the target hand in the to-be-processed image.

Images in the video stream and the to-be-processed image are obtained by capturing a same target area, and the image in the video stream is obtained by capturing subsequent to capture of the to-be-processed image.

In some embodiments, the to-be-processed image and the image in the video stream may be synchronously captured.

In some embodiments, the to-be-processed image is an image of the target area captured at an overhead view, while the image in the video stream may be an image of the target area captured at an overhead view or an image of the target area captured at other views, for example, at a side view.

In response to determining the target hand from the to-be-processed image, the target hand is tracked from the synchronously captured video stream.

For the target hand in the to-be-processed image, a hand associated with the target hand is determined from the image in the video stream, and the hand is determined as a target hand in the video stream for tracking. An association relationship between the hand in the to-be-processed image and the hand in the image in the video stream may be determined by using various ways, and no limitation is made thereto in the present disclosure.

In the present disclosure, by detecting a hand pose in a bounding box of hand in a hand detection result to determine a target hand that conforms to a hand pose in a target gesture, i.e., recognizing a target hand to be tracked from a to-be-processed image, and tracking the target hand in a synchronously captured video stream, the target hand can be recognized quickly and effectively according to the target gesture, thereby improving tracking efficiency.

In some embodiments, a center point of the bounding box of hand as the position of the hand.

A person skilled in the art should understand that other position points of the bounding box of hand may also be determined as the position of the hand, for example, when a target gesture to be recognized has a closest relationship with one of fingers, the position of the finger in the bounding box of hand may be determined as the position of the hand. The specific determining mode of the position of the hand may be determined according to the relationship between the target gesture to be recognized and the bounding box of hand.

In some embodiments, in response to the hand detection result including a bounding box of hand, it is determined a hand in the bounding box of hand, with the hand pose conforming to the hand pose in the target gesture and in a pre-defined area in the to-be-processed image, as the target hand.

In the embodiments of the present disclosure, in a case that the hand detection result includes a bounding box of hand, the hand detection result also includes position information of the bounding box of hand, i.e., position information of the hand in the to-be-processed image is determined. For example, the center point position of the bounding box of hand may be taken as the position of the hand. In a case that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture, the hand in the pre-defined area in the to-be-processed image, where the center point of the bounding box of hand is located, may be taken as the target hand.

In one example, the pre-defined area in the to-be-processed image includes two set sub-areas corresponding to two sub-areas in the target area respectively in the to-be-processed image. For example, in a case that the target area is a game table, the pre-defined area includes two set sub-areas corresponding to two sub-areas on the game table.

For a specific hand pose occurring in a specific area referred to as the target gesture, a position constraint is provided for the hand in the bounding box of hand by using the pre-defined area in the to-be-processed image to further determine that the detected hand pose is the target gesture. The pre-defined area in the to-be-processed image may be determined according to the position and intrinsic parameters of the camera for capturing the to-be-processed image, and the specific area corresponding to the target gesture in an actual space.

In some embodiments, it may be determined that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture in the following way: inputting the image in the bounding box of hand to a pre-trained hand classification neural network to obtain a hand classification result, and determining whether the hand pose in the bounding box of hand conforms to the hand pose in the target gesture by means of the classification result.

The used hand classification neural network may be a convolutional classification model, for example, a convolutional classification model using Rsenet 18 as a backbone network. A person skilled in the art should understand that the used hand classification neural network may be determined according to a classification type to be obtained; the classification type to be obtained may be determined according to a target gesture to be recognized. No limitation is made thereto in the present disclosure.

In one example, when the target gesture to be recognized relates only to a hand facing up, the classification type to be obtained may include the hand facing up or the hand with other poses, where the other poses include hand poses other than the hand facing up and the hand being blocked. The hand classification neural network may be trained with an image sample including the hand facing up, and an annotation result of the image sample with the hand facing up is the palm of the hand.

In one example, when the target gesture to be recognized relate to the hand facing up and the hand facing back, the classification type to be obtained may include the hand facing up, the hand facing back, and the hand with other poses, where the other poses include hand poses other than the hand facing up and the hand facing back, and the hand being blocked. The hand classification neural network may be trained with the image sample including the hand facing up and an image sample including the hand facing back, where the annotation result of the image sample with the hand facing up is the palm of the hand, and an annotation result of the image sample with the hand facing back may be an identifier indicating the back.

In one example, when the target gesture to be recognized relates to a specific pose, the classification type needs to include the specific pose, and then the hand classification neural network may be trained with a sample hand image including the pose, so that the hand classification neural network can recognize the specific pose from the image in the bounding box of hand.

A person skilled in the art should understand that it may also be determined that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture in other ways, gesture detection is performed by using the hand pose in the bounding box of hand, and the like. No limitation is made thereto in the present disclosure.

In the embodiments of the present disclosure, images in the bounding box of hand are classified with the hand classification neural network, so as to recognize whether the pose is the hand facing up, i.e., the pose where the palm can be seen. In this case, for the used hand classification neural network, compared with a video classification mode, it is easy to obtain and

annotate a training sample, the recognition speed of the model is high, and thus, real-time recognition for the hand pose can be implemented.

In some embodiments, the target hand may be tracked in a video stream according to the target hand in the to-be-processed image in the following way.

For an image frame in the video stream, a first matrix and a second matrix of the image frame are determined according to a first matrix and a second matrix of a previous image frame of the image frame, where the first matrix of the image frame includes position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

In response to the target hand being recognized from the to-be-processed image, i.e., the target hand to be tracked is recognized from the to-be-processed image, the target hand to be tracked may be tracked in the video stream captured at a same view or in the video stream captured at different views.

When the target hand to be tracked is tracked in the video stream captured at different views, it is required to first determine the hand associated with the recognized target hand from the image captured at the same view as the video stream and captured synchronously with the to-be-processed image as the target hand to be tracked, and then the target hand to be tracked is tracked in the video stream.

How to track the target hand to be tracked in the video stream is described below by taking the to-be-processed image and the video stream captured at the same view as an example.

For the video stream captured subsequent to capture of the to-be-processed image, a first matrix and a second matrix of each image frame are predicted according to a first matrix and a second matrix of a previous image frame. Initial information of a first matrix X and a second matrix P may be determined according to the hand detection result of the to-be-processed image. The first matrix X includes position information of the center of the bounding box of hand of the target hand in the image frame, for example, the matrix may be 1*2 matrix; the second matrix P measures each element in the first matrix X, i.e., a covariance matrix of total errors of position variables, and the matrix is an artificially set hyperparameter matrix, for example, the matrix may be a 2*2 matrix.

For any image frame in the video stream, in a case of different hand detection results, the first matrix and the second matrix of the image frame may be determined in different ways.

In response to the hand detection result of the image frame not including the bounding box of

hand of the target hand, it is assumed that the target hand is in a stationary state, the first matrix X of the previous image frame is determined as the first matrix X in the image frame, i.e., the first matrix X remains unchanged; for the second matrix P in the image frame, the second matrix in the previous image frame is corrected with a correction matrix Q to obtain the second matrix of the image frame, i.e., the second matrix in the image frame is determined as the second matrix of the previous image frame + the correction matrix (P+Q), where the correction matrix Q represents an error of each element in the second matrix of the previous image frame, i.e., being used for measuring an error of the second matrix P, the matrix is also an artificially set hyperparameter matrix and is a matrix having a same size as that of the second matrix P. In a case that the second matrix P is a 2*2 matrix, the correction matrix Q is also a 2*2 matrix.

In response to the hand detection result of the image frame including the bounding box of hand of the target hand, a result of prediction according to the previous image frame is corrected according to the hand detection result to obtain the first matrix X and the second matrix P of the image frame.

First, the first matrix X and the second matrix P of the image frame may be predicted according to the first matrix X of the previous image frame according to a time interval between the image frame and the previous image frame. For example, it may be assumed that the target hand does uniform rectilinear motion in a set direction, and on the basis that the first matrix X including the position information of the previous image frame is known, the first matrix X and the second matrix P of the image frame may be determined according to the time interval between the image frame and the previous image frame. The first matrix X and the second matrix P here may be referred to as a first prediction matrix and a second prediction matrix.

Next, a third matrix O of the image frame may be obtained according to the position information of the bounding box of hand of the target hand, and a fourth matrix R that represents an error of each element in the third matrix is obtained. The fourth matrix R measures each element in the third matrix O, i.e., a covariance matrix of total errors of position variables, and the matrix is also an artificially set hyperparameter matrix. The third matrix O and the fourth matrix R here may be referred to as a third prediction matrix and a fourth prediction matrix.

According to the hand detection result (such as the position information of the bounding box of hand of the target hand) of the image frame, the determined fourth matrix R of the image frame and the predicted second matrix P both comply with Gaussian distributions, new Gaussian distributions may be obtained by multiplying Gaussian distributions corresponding to the fourth matrix R by Gaussian distributions corresponding to the second matrix P, a new matrix X may be obtained according to an average value of the obtained new Gaussian distributions, and a new

matrix P may be obtained according to covariance of the obtained new Gaussian distributions. The previously predicted first matrix X and second matrix P are updated with the obtained new first matrix X and new second matrix P, and then the first matrix X including the position information in the image frame may be obtained, so as to implement tracking of the target hand in the video stream.

In the embodiments of the present disclosure, the position information of the target hand in the previous image frame is obtained with the matrix and covariance matrix including the position information of the target hand of the previous image frame, and thus, the tracking of the target hand in the video stream can be implemented accurately and effectively.

The target tracking method provided by the embodiments of the present disclosure is described below by taking the target gesture to be recognized being two hands both facing up and the two hands being respectively located above two specific sub-areas on a tabletop as an example. A person skilled in the art should understand that the target tracking method is not limited to tracking the target hand making the target gesture above, and may also be used for tracking the target hand making other target gestures.

FIG. 2 is a schematic diagram of an application scene of a target tracking method provided by at least one embodiment of the present disclosure. As shown in FIG. 2, the target gesture to be recognized is the gesture with two palms facing up made above two areas 221 and 222 of a tabletop 20.

First, a to-be-processed image including all or some of areas of the tabletop 20 captured at an overhead view is detected so as to obtain a hand detection result.

If a hand is detected, i.e., the hand detection result includes a bounding box of hand, an image in the bounding box of hand is input to a hand classification neural network to obtain a hand classification result. The hand classification neural network is used for classifying the input image with the hand facing up or the hand with other poses, and an obtained hand classification result including the hand facing up or the hand with the other poses. If classification results of two bounding boxes of hand are hands facing up, it is determined that hand poses in the bounding boxes of hand are both the hands facing up and conform to hand poses in the target gesture.

If it is determined that the two hands both face up, it is determined whether the positions of the two bounding boxes of hand are located in two set sub-areas of the to-be-processed image. The two set sub-areas of the to-be-processed image are determined according to the position and intrinsic parameters of the camera for capturing the to-be-processed image, and the positions of two areas 221 and 222 on the tabletop 20. In response to center points of the two bounding boxes

of hand are respectively located in the two set sub-areas, it is determined that the hands make the target gesture.

In one example, an image cropped according to the bounding box of hand may be input into the hand classification neural network to obtain a hand classification result. If the to-be-processed image includes multiple bounding boxes of hand, multiple cropped images are input into the hand classification neural network respectively.

The above hand classification neural network used for classifying the input image with a hand facing up or the hand with other poses may be trained in the following way.

First, an image sample including a hand facing up is obtained, and the palm of the hand in the image sample is annotated.

FIGS. 3A and 3B are exemplary image samples with hands facing up. The hand image with the palm facing up may be taken as an image sample with the hand facing up, and an annotation result of the image sample of the hand facing up is the palm of the hand. White points in the palms in FIGS. 3A and 3B are annotations of the image samples with the hands facing up. FIGS. 3C-3F are exemplary hand images with other poses which are a hand image with a hand facing back or a palm being blocked, and the hand images may be annotated with other poses.

The hand classification neural network is trained with the image sample above, and the trained neural network can classify the image in the bounding box of hand to obtain a classification result with the hand facing up or the hand with other poses.

In some embodiments, the hand classification neural network may classify the image in the bounding box of hand in three ways to obtain a classification result with the hand facing up, or the hand facing back, or the hand with other poses. The hand classification neural network may be trained by obtaining an image sample with the hand facing up, the hand facing back, and the hand with other poses. FIGS. 3A-3F as still taken as examples. FIGS. 3A and 3B are exemplary image samples with hands facing up, and an annotation result of the image sample with the hand facing up is the palm of the hand. FIGS. 3C and 3D are exemplary image samples with hands facing back. The hand image with a complete hand back and five fingers seen may be taken as a back sample hand image, and the hand image is annotated with a back. FIGS. 3E and 3F are exemplary image samples with hands with other poses. The hand image with a palm, a complete hand back, or five fingers not seen may be taken as an image sample with the hand with other poses, and the hand image is annotated with other poses. In FIGS. 3E and 3F, although the overall poses face up, palms cannot be seen, and therefore, the poses are other poses. The hand classification neural network is trained with the image sample above, and the trained neural

network can classify the image in the bounding box of hand to obtain a classification result with the hand facing up, the hand facing back, or the hand with other poses.

In the embodiments of the present disclosure, the image in the bounding box of hand is classified with the hand classification neural network, so as to recognize whether the hand presents a pose facing up, i.e., the pose with the palm facing up. In this case, for the used hand classification neural network, compared with a video classification mode, it is easy to obtain and annotate a training sample, the recognition speed of the model is high, and thus, real-time recognition for the hand pose can be implemented.

On the basis that the hand classification result conforms to the hand pose in the target gesture, a position constraint is performed on a pre-defined area of the to-be-processed image to further determine that the detected hand makes the target gesture.

In response to a target hand making a gesture with two palms facing up above two areas 221 and 222 of a tabletop 20 is recognized from the to-be-processed image including the tabletop 20, the target hand is tracked from a synchronously captured video stream. For example, position information of the target hand in a current image frame is obtained with a matrix and covariance matrix including the position information of the target hand of a previous image frame, and thus, the tracking of the target hand in the video stream can be implemented accurately and effectively.

The target tracking method provided by the embodiments of the present disclosure is described by taking the target gesture being a palm-up gesture of a croupier as an example.

FIG. 4 is a schematic diagram of an application scene of a target tracking method provided by at least one embodiment of the present disclosure. As shown in FIG. 4, in a game scene, a game may be played by means of a game table 40, and players 421, 422, 423 and a croupier 43 are located on opposite sides of the game table. The game table mainly includes: an exchanged object area 47 (an area in front of the croupier) of the croupier, exchanged object areas 441 and 442 (areas in front of the players) of the players, a game area 46, and a pre-defined area 41 (which can also be referred to as a hot area).

The palm-up gesture of the croupier indicates that after completing a specific operation or before executing the specific operation, the palms of both hands of the croupier face up and the hands are horizontally placed on the two sides of the exchanged object area of the croupier (as the exchanged object area 47 in FIG. 4) to show others that there is nothing in the hands, indicating that the game can proceed to a next stage.

In the embodiments of the present disclosure, two sub-areas on the two sides of the exchanged object area in front of the croupier on the game table (for example, sub-areas 411 and 412 in FIG.

4) may be taken as a specific area of an actual space. Two corresponding set sub-areas on a to-be-processed image are determined according to the position and intrinsic parameters of a camera for capturing the to-be-processed image, and the position of the specific area, and a position constraint is performed on a hand classification result.

The to-be-processed image including all or some of areas of the game table 40 is detected. If an obtained hand detection result includes a bounding box of hand, an image in the bounding box of hand is input into a pre-trained hand classification neural network. The hand classification neural network is capable of classifying the input image with a hand facing up or the hand with other poses. In response to classification results of images in two bounding boxes of hand being both hands facing up, it is further determined whether two hand detection positions are located in the two set sub-areas in the to-be-processed image. If the two hands are detected to be located in the two set sub-areas respectively, it is indicated that the two hands enter the hot area on the game table 40 respectively, and then it can be determined that the detected hands make the palm-up gesture.

In response to the target hand making the palm-up gesture being recognized from the to-be-processed image including the game table 40, i.e., the hands of the croupier are recognized, the hands of the croupier are tracked from a synchronously captured video stream. For example, position information of the hands of the croupier in a current image frame is obtained with a matrix and covariance matrix including position information of the target hand of a previous image frame, and thus, the tracking of the hands of the croupier in the video stream can be implemented accurately and effectively. By recognizing the hand of the croupier, whether any forbidden objects are carried in the hands of the croupier can be monitored.

FIG. 5 is a schematic structural diagram of a target tracking apparatus provided by at least one embodiment of the present disclosure. As shown in FIG. 5, the apparatus includes: a detecting unit 501, configured to detect a to-be-processed image to obtain a hand detection result; a determining unit 502, configured to, in response to the hand detection result including a bounding box of hand, determine a hand with a hand pose in the bounding box of hand conforming to a hand pose in a target gesture as a target hand; and a tracking unit 503, configured to track the target hand in a video stream according to the target hand in the to-be-processed image, where images in the video stream and the to-be-processed image are obtained by capturing a same target area, and the image in the video stream is obtained by capturing subsequent to the to-be-processed image.

In some embodiments, the determining unit 502 is specifically configured to: in response to the hand detection result including the bounding box of hand, determining a hand in the bounding box of hand, with the hand pose conforming to the hand pose in the target gesture and in a pre-

defined area in the to-be-processed image, as the target hand.

In some embodiments, the determining unit 502 is specifically configured to, in response to the hand detection result including the bounding box of hand, crop an image located in the bounding box of hand in the to-be-processed image; and determine, according to the cropped image, that the hand pose in the cropped image conforms to the hand pose in the target gesture.

In some embodiments, the hand pose in the target gesture is that two hands both face up; the determining unit 502 is configured to: input the image in the bounding box of hand to a hand classification neural network to obtain a hand classification result, the hand classification result including a hand facing up or the hand with other poses; and if classification results in two bounding boxes of hand are that two hands both face up, determine that the hand poses in the bounding boxes of hand conform to the hand pose in the target gesture.

In some embodiments, the hand classification neural network is obtained by training with hand image samples, a hand image sample including an image sample with the hand facing up, and an annotation result of the image sample with the hand facing up being the palm of the hand.

In some embodiments, the determining unit 502 configured to determine that the hand in the bounding box of hand is located in the pre-defined area in the to-be-processed image, is configured to: determine that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image.

In some embodiments, the pre-defined area includes two set sub-areas corresponding to two sub-areas in the target area respectively in the to-be-processed image; the determining unit 502 configured to determine that the center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image, is configured to determine that the center points of the two bounding boxes of hand are located in the two set sub-areas respectively.

In dome embodiments, the tracking unit 503 is configured to: for an image frame in the video stream, a first matrix and a second matrix of the image frame are determined according to a first matrix and a second matrix of a previous image frame of the image frame, where the first matrix of the image frame includes position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

In some embodiments, the tracking unit 503 configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is configured to: in response to a hand detection result of the image frame not including the bounding box of hand of the target hand, determine the first

matrix of the previous image frame as the first matrix of the image frame, and correct the second matrix of the previous image frame with a correction matrix to obtain the second matrix of the image frame, where the correction matrix represents an error of each element in the second matrix of the previous image frame.

In some embodiments, the tracking unit 503 configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is configured to: in response to the hand detection result of the image frame including the bounding box of hand of the target hand, predict, according to a time interval between the image frame and the previous image frame, the first matrix and the second matrix of the image frame according to the first matrix of the previous image frame; obtain a third matrix and a fourth matrix of the image frame according to the position information of the bounding box of hand of the target hand, where the third matrix includes the position information of the center of the bounding box of hand of the target hand in the image frame, and the fourth matrix represents an error of each element in the third matrix; and multiply Gaussian distributions corresponding to the second matrix by Gaussian distributions corresponding to the fourth matrix to obtain new Gaussian distributions, update the first matrix according to an average value of the obtained new Gaussian distributions, and update the second matrix according to covariance of the obtained new Gaussian distributions. In some embodiments, the to-be-processed image is an image of the target area captured at an overhead view.

FIG. 6 is a schematic structural diagram of an electronic device provided according to at least one embodiment of the present disclosure. The device includes: a processor; and a memory configured to store processor-executable instructions, where the processor is configured to invoke the instructions stored in the memory to execute the target tracking method according to at least one embodiment.

A person skilled in the art should understand that one or more embodiments of the description may provide a method, a system or a computer program product. Therefore, one or more embodiments of the description may take the forms of hardware embodiments, software embodiments, or embodiments in combination with software and hardware. Moreover, one or more embodiments of the description may use the form of the computer program product implemented over one or more computer usable storage media (including but not limited to a disk memory, a CD-ROM, and an optical memory, etc.) that include a computer usable program code.

The embodiments of the description further provide a computer-readable storage medium, having a computer program stored thereon, where the program is executed by a processor, steps of the target tracking method described in any one of embodiments of the present disclosure is

implemented.

The embodiments in the description are all described in a progressive manner, for same or similar parts in the embodiments, refer to these embodiments, and each embodiment focuses on a difference from other embodiments. In particular, data processing device embodiments are substantially similar to method embodiments and therefore are only described briefly, and for the associated part, refer to the descriptions of the method embodiments.

The specific embodiments of the description are described above. Other embodiments are within the scope of the appended claims. In some cases, actions or steps described in the claims may be performed in an order different from that in the embodiments and can still achieve a desired result. In addition, the processes described in the accompanying drawings do not necessarily require a specific order shown or a sequential order to achieve the desired result. In some implementations, multi-task processing and parallel processing may also be performed or may be advantageous.

The embodiments of the subject matter and functional operations described in the description may be implemented in digital electronic circuitry, tangible computer software or firmware, computer hardware including the structures disclosed in the description and structural equivalents thereof, or a combination of one or more thereof. The embodiments of the subject matter described in the description may be implemented as one or more computer programs, i.e., one or more modules of computer program instructions encoded on a tangible non-transitory program carrier to be executed by a data processing apparatus or to control operations of the data processing apparatus. Alternatively or additionally, the program instructions may be encoded on artificially generated propagated signals, such as machine-generated electrical, optical or electromagnetic signals, generated to encode and transmit information to a suitable receiver apparatus for execution by the data processing apparatus. The computer storage medium may be a machine-readable storage device, a machine-readable storage substrate, a random or serial access memory device, or a combination of one or more thereof.

The processes and logic flows described in the description can be performed by one or more programmable computers executing one or more computer programs to perform corresponding functions by performing operations according to input data and generating output. The processes and logic flows may also be performed by a special logic circuit, such as a Field Programmable Gate Array (FPGA) or an Application Specific Integrated Circuit (ASIC), and the apparatus may also be implemented as a special logic circuit.

The computer suitable for executing the computer program includes, for example, a general-

purpose microprocessor and/or a special-purpose microprocessor, any other type of central processing unit. Generally, the central processing unit receives instructions and data from a read-only memory and/or a random access memory. Basic components of the computer include a central processing unit for implementing or executing instructions and one or more memory devices for storing instructions and data. Generally, the computer further includes one or more large-capacity storage devices for storing data, for example, a magnetic disk, a magneto-optical disk, or an optical disk, or the computer is operably coupled to the large-capacity storage device to receives data therefrom or transmit data thereto, or receive data therefrom and transmit data therefrom. However, the computer does not necessarily include such a device. Furthermore, the computer may be embedded in another device, for example, a mobile phone, a Personal Digital Assistant (PDA), a mobile audio or a video player, a game console, a Global Positioning System (GPS) receiver, or a portable storage device, for example, a Universal Serial Bus (USB) flash drive, just a few examples provided.

A computer-readable medium suitable for storing computer program instructions and data include a non-volatile memory, a medium, and a memory device in all forms, including, for example, a semiconductor memory device (for example, an EPROM, an EEPROM, and a flash device), a magnetic disk (for example, an internal hardware or a movable disk), a magneto-optical disk, and a CD ROM and DVD-ROM disk. The processor and the memory may be supplemented by the special logic circuit or incorporated into the special logic circuit

Although the description includes many specific implementation details, these should not be interpreted as limiting the scope of any invention or the scope of protection, and are mainly used for describing the features of specific embodiments of a specific invention. Some features described in multiple embodiments in the description may also be implemented in combination in a single embodiment. In addition, various features described in a single embodiment may be separately implemented in multiple embodiments or in any suitable sub-combination. Furthermore, although the features may function in some combinations as described above and even set forth in such a way initially, one or more features from a claimed combination may be removed from the combination in some cases, and the claimed combination may relate to a sub-combination or a modification of the sub-combination.

Similarly, although operations are described in the accompanying drawings in a specific order, this should not be understood as requiring that such operations are performed in the specific order shown or in a sequential order, or that all illustrated operations are performed to achieve a desired result. In some cases, multi-task and parallel processing may be advantageous. Furthermore, the separation of various system modules and components in the embodiments above should not be

understood as requiring such separation in all embodiments, and it should be understood that the described program components and systems can generally be integrated together in a single software product or packaged into multiple software products.

Thus, specific embodiments of the subject matter have been described. Other embodiments are within the scope of the appended claims. In some cases, the actions described in the claims can be performed in a different order and still achieve the desired result. In addition, the processes described in the accompanying drawings do not necessarily require a specific order shown or a sequential order to achieve the desired result. In some cases, multi-task and parallel processing may be advantageous.

The above descriptions are only preferred embodiments of one or more embodiments of the description and are not intended to limit one or more embodiments of the description. Any modifications, equivalent substitutions and improvements made without departing from the spirit and principle of one or more embodiments of the description are intended to be included within the scope of one or more embodiments of the description.

CLAIMS

1. A target tracking method, comprising:

detecting a to-be-processed image to obtain a hand detection result;

in response to the hand detection result comprising a bounding box of hand, determining a hand in the bounding box of hand, with a hand pose conforming to a hand pose in a target gesture and in a pre-defined area in the to-be-processed image, as a target hand; and

tracking the target hand in a video stream according to the target hand in the to-be-processed image, wherein images in the video stream and the to-be-processed image are obtained by capturing a same target area, the images in the video stream are obtained by capturing subsequent to capture of the to-be-processed image, the target area comprises two sub-areas, the pre-defined area in the to-be-processed image comprises two set sub-areas corresponding to the two sub-areas respectively, the hand in the bounding box in the pre-defined area refers to that the hand in the bounding box is located in any of the two set sub-areas, and the pre-defined area in the to-be-processed image is determined based on the target area.

2. The method according to claim 1, wherein determining that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture comprises:

in response to the hand detection result comprising the bounding box of hand, cropping an image located in the bounding box of hand in the to-be-processed image; and

determining, according to the cropped image, that a hand pose in the cropped image conforms to the hand pose in the target gesture.

3. The method according to any one of claims 1 or 2, wherein the hand pose in the target gesture is that two hands both face up;

determining that the hand pose in the bounding box of hand conforms to the hand pose in the target gesture comprises:

inputting the image in the bounding box of hand to a hand classification neural network to obtain a hand classification result, the hand classification result comprising a hand facing up or the hand with other poses; and

if classification results in two bounding boxes of hand are that two hands both face up and

two hands in the two bounding boxes are respectively located in the two set sub-areas, determining that the hand poses in the two bounding boxes of hand conform to the hand pose in the target gesture.

4. The method according to claim 3, wherein the hand classification neural network is obtained by training with hand image samples, a hand image sample comprising an image sample with the hand facing up, and an annotation result of the image sample with the hand facing up being the palm of the hand.

5. The method according to claim 1, wherein determining that the hand in the bounding box of hand is located in the pre-defined area in the to-be-processed image comprises:

determining that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image.

6. The method according to any one of claims 1 to 5, wherein the tracking the target hand in a video stream according to the target hand in the to-be-processed image comprises:

for an image frame in the video stream, determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame,

wherein the first matrix of the image frame comprises position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

7. The method according to claim 6, wherein the determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame comprises:

in response to a hand detection result of the image frame not comprising the bounding box of hand of the target hand, determining the first matrix of the previous image frame as the first matrix of the image frame, and correcting the second matrix of the previous image frame with a correction matrix to obtain the second matrix of the image frame, wherein the correction matrix represents an error of each element in the second matrix of the previous image frame.

8. The method according to claim 6, wherein the determining a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame comprises:

in response to the hand detection result of the image frame comprising the bounding box of hand of the target hand, predicting, according to a time interval between the image frame and the previous image frame, the first matrix and the second matrix of the image frame according to the first matrix of the previous image frame;

obtaining a third matrix and a fourth matrix of the image frame according to the position information of the bounding box of hand of the target hand, wherein the third matrix comprises the position information of the center of the bounding box of hand of the target hand in the image frame, and the fourth matrix represents an error of each element in the third matrix; and

multiplying Gaussian distributions corresponding to the second matrix by Gaussian distributions corresponding to the fourth matrix to obtain new Gaussian distributions, updating the first matrix according to an average value of the obtained new Gaussian distributions, and updating the second matrix according to covariance of the obtained new Gaussian distributions.

9. A target tracking apparatus, comprising:

a detecting unit, configured to detect a to-be-processed image to obtain a hand detection result;

a determining unit, configured to: in response to the hand detection result comprising a bounding box of hand, determine in the bounding box of hand, with a hand pose conforming to a hand pose in a target gesture and in a pre-defined area in the to-be-processed image, as a target hand; and

a tracking unit, configured to track the target hand in a video stream according to the target hand in the to-be-processed image, wherein images in the video stream and the to-be-processed image are obtained by capturing a same target area, the images in the video stream are obtained by capturing subsequent to capture of the to-be-processed image, the target area comprises two sub-areas, the pre-defined area in the to-be-processed image comprises two set sub-areas corresponding to the two sub-areas respectively, the hand in the bounding box in the pre-defined area refers to that the hand in the bounding box is located in any of the two set sub-areas, and the pre-defined area in the to-be-processed image is determined based on the target area.

10. The apparatus according to claim 9, wherein the determining unit is configured to:

in response to the hand detection result comprising the bounding box of hand, crop an image located in the bounding box of hand in the to-be-processed image; and

determine, according to the cropped image, that a hand pose in the cropped image conforms to the hand pose in the target gesture.

11. The apparatus according to any one of claims 9 or 10, wherein the hand pose in the target gesture is that two hands both face up;

the determining unit is configured to:

input the image in the bounding box of hand to a hand classification neural network to obtain a hand classification result, the hand classification result comprising a hand facing up or the hand with other poses; and

if classification results in two bounding boxes of hand are that two hands both face up and two hands in the two bounding boxes are respectively located in the two set sub-areas, determine that the hand poses in the two bounding boxes of hand conform to the hand pose in the target gesture.

12. The apparatus according to claim 11, wherein the hand classification neural network is obtained by training with hand image samples, a hand image sample comprising an image sample with the hand facing up, and an annotation result of the image sample with the hand facing up being the palm of the hand.

13. The apparatus according to claim 9, wherein the determining unit configured to determine that the hand in the bounding box of hand is located in the pre-defined area in the to-be-processed image, is configured to:

determine that a center point of the bounding box of hand is located in the pre-defined area of the to-be-processed image.

14. The apparatus according to any one of claims 9 to 13, wherein the tracking unit is

configured to:

for an image frame in the video stream, determine a first matrix and a second matrix of the image frame according to a first matrix and a second matrix of a previous image frame of the image frame,

wherein the first matrix of the image frame comprises position information of the center of the bounding box of hand of the target hand in the image frame, and the second matrix of the image frame represents an error of each element in the first matrix of the image frame.

15. The apparatus according to claim 14, wherein the tracking unit configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is configured to:

in response to a hand detection result of the image frame not comprising the bounding box of hand of the target hand, determine the first matrix of the previous image frame as the first matrix of the image frame, and correct the second matrix of the previous image frame with a correction matrix to obtain the second matrix of the image frame, wherein the correction matrix represents an error of each element in the second matrix of the previous image frame.

16. The apparatus according to claim 14, wherein the tracking unit configured to determine the first matrix and the second matrix of the image frame according to the first matrix and the second matrix of the previous image frame of the image frame, is specifically configured to:

in response to the hand detection result of the image frame comprising the bounding box of hand of the target hand, predict, according to a time interval between the image frame and the previous image frame, the first matrix and the second matrix of the image frame according to the first matrix of the previous image frame;

obtain a third matrix and a fourth matrix of the image frame according to the position information of the bounding box of hand of the target hand, wherein the third matrix comprises the position information of the center of the bounding box of hand of the target hand in the image frame, and the fourth matrix represents an error of each element in the third matrix; and

multiply Gaussian distributions corresponding to the second matrix by Gaussian distributions corresponding to the fourth matrix to obtain new Gaussian distributions, update the first matrix according to an average value of the obtained new Gaussian distributions, and update the second

matrix according to covariance of the obtained new Gaussian distributions.

17. An electronic device, comprising a memory and a processor, wherein the memory is configured to store computer instructions runnable on the processor, and the processor is configured to implement the method according to any one of claims 1 to 8 when executing the computer instructions.

18. A computer-readable storage medium, having a computer program stored thereon, wherein when the program is executed by a processor, the method according to any one of claims 1 to 8 is implemented.

19. A method of tracking a hand of croupier, comprising:

detecting a to-be-processed image comprising some or all of areas of a game table, to obtain a hand detection result;

determining whether a hand in each of two bounding boxes of hand comprised in the hand detection result faces up, and whether two hands in the two bounding boxes are respectively located in two set sub-areas of the to-be-processed image, wherein the game table comprises two sub-areas, the two set sub-areas corresponds to the two sub-areas respectively and are determined based on the two sub-areas;

when determining that the hand in each of the two bounding boxes faces up and that the two hands in the two bounding boxes are respectively located in the two set sub-areas of the to-be-processed image, determining that the two hands of the croupier make a palm-up gesture; and

tracking the two hands of the croupier in a video stream, wherein images in the video stream and the to-be-processed image are obtained by capturing the game table, the images in the video stream are obtained by capturing subsequent to capture of the to-be-processed image.