

OBJECT STATE RECOGNITION METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

[0001] This application claims priority to Chinese Patent Application No. 2024109290569, entitled "OBJECT STATE RECOGNITION METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM" filed with the China National Intellectual Property Administration on July 11, 2024, which is incorporated herein by reference in its entirety.

FIELD OF THE TECHNOLOGY

[0002] This application relates to the field of computer technologies, and in particular, to an object state recognition method and apparatus, a computer device, a storage medium, and a computer program product.

BACKGROUND OF THE DISCLOSURE

[0003] With the development of computer technologies, an Internet of Things technology appears. The Internet of Things refers to connecting any item to a network through information sensing devices according to agreed protocols. Items exchange information and communicate with each other through information transmission media, to implement functions such as intelligent recognition, positioning, tracking, and supervision.

[0004] Currently, a user state can be recognized by using the Internet of Things technology. A motion change and a motion direction of a wearable device in space are mainly measured by using sensors such as an accelerometer, a gyroscope, and a magnetometer in the wearable device, to determine a body posture and a motion direction of a wearer, so as to recognize an object state of the wearer. However, some object states inherently exhibit similarity in kinematic signal representation, and are difficult to be distinguished from each other. A current recognition method has a problem of low accuracy.

SUMMARY

[0005] Embodiments of this application provide an object state recognition method and apparatus, a computer device, a computer-readable storage medium, and a computer program product.

- [0006] This application provides an object state recognition method. The method is performed by a computer device and includes:
- [0007] obtaining a target video;
- [0008] obtaining a background frame corresponding to the target video, and determining, from the target video, a target frame corresponding to a target action of a target object;
- [0009] performing image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame;
- [0010] generating an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item; and
- [0011] determining, based on each interaction information group, a target object state corresponding to the target object in the target frame.
- [0012] This application further provides an object state recognition apparatus. The apparatus includes:
- [0013] a video obtaining module, configured to obtain a target video;
- [0014] an image determining module, configured to: obtain a background frame corresponding to the target video, and determine, from the target video, a target frame corresponding to a target action of a target object;
- [0015] an image segmentation module, configured to perform image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame;
- [0016] an information group generation module, configured to generate an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item; and
- [0017] a state recognition module, configured to determine, based on each interaction information group, a target object state corresponding to the target object in the target frame.
- [0018] This application further provides a computer device. The computer device includes a memory and a processor. The memory stores computer-readable instructions, and the processor,

when executing the computer-readable instructions, implements operations of the foregoing object state recognition method.

[0019] This application further provides a computer-readable storage medium. The computer-readable storage medium has computer-readable instructions stored therein. The computer-readable instructions, when executed by a processor, implement operations of the foregoing object state recognition method.

[0020] This application further provides a computer program product. The computer program product includes computer-readable instructions. The computer-readable instructions, when executed by a processor, implement operations of the foregoing object state recognition method.

[0021] Details of one or more embodiments of this application are provided in the accompany drawings and descriptions below. Other features, objectives, and advantages of this application will become apparent in the specification, the accompanying drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] To describe the technical solutions in the embodiments of this application or in the related technology more clearly, the following briefly describes the accompanying drawings required for describing the embodiments or the related technology. Apparently, the accompanying drawings in the following descriptions show merely some embodiments of this application, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

[0023] FIG. 1 is a diagram of an application environment of an object state recognition method according to an embodiment.

[0024] FIG. 2 is a schematic flowchart of an object state recognition method according to an embodiment.

[0025] FIG. 3 is a schematic diagram of determining a background frame from a target video according to an embodiment.

[0026] FIG. 4 is a schematic flowchart of determining a target frame from a target video according to an embodiment.

[0027] FIG. 5 is a schematic diagram of recognizing a target frame according to an embodiment.

[0028] FIG. 6 is a schematic diagram of performing interaction analysis based on an image segmentation result according to an embodiment.

[0029] FIG. 7 is a schematic flowchart of determining a target object state from a reference object state according to an embodiment.

[0030] FIG. 8 is a schematic flowchart of determining a target object state through an interaction model according to an embodiment.

[0031] FIG. 9 is a schematic diagram of a falling detection method according to an embodiment.

[0032] FIG. 10 is a schematic diagram of a falling detection method according to another embodiment.

[0033] FIG. 11 is a structural block diagram of an object state recognition apparatus according to an embodiment.

[0034] FIG. 12 is a diagram of an internal structure of a computer device according to an embodiment.

[0035] FIG. 13 is a diagram of an internal structure of a computer device according to another embodiment.

DESCRIPTION OF EMBODIMENTS

[0036] To make the objectives, technical solutions, and advantages of this application clearer, the following further describes this application in detail with reference to the accompanying drawings and the embodiments. The specific embodiments described herein are merely configured for explaining this application but are not intended to limit this application.

[0037] An object state recognition method provided in the embodiments of this application may be applied to an application environment shown in FIG. 1. A terminal 102 communicates with a server 104 through a network. A data storage system can store data that needs to be processed by the server 104. The data storage system may be independently disposed, may be integrated on the server 104, or may be disposed on a cloud or another device. The terminal 102 may be, but is not limited to, a personal computer, a notebook computer, a smartphone, a tablet computer, an Internet of Things device, and a portable wearable device. The Internet of Things device may be a smart speaker, a smart television, a smart air conditioner, a smart vehicle-mounted device, or the like. The portable wearable device may be a smart watch, a smart bracelet, a head-mounted device, or the like.

The server 104 may be implemented by using an independent server, a server cluster including a plurality of servers, or a cloud server.

[0038] Both the terminal and the server each may be independently configured to perform the object state recognition method provided in the embodiments of this application. Alternatively, the terminal and the server may be collaboratively configured to perform the object state recognition method provided in the embodiments of this application.

[0039] For example, the server obtains a target video, obtains a background frame corresponding to the target video, and determines, from the target video, a target frame corresponding to a target action of a target object. The server performs image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame. The server generates an interaction information group based on position information of the target region and position information of the item region, an item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item; and determines, based on each interaction information group, a target object state corresponding to the target object in the target frame.

[0040] In an embodiment, as shown in FIG. 2, an object state recognition method is provided. An example in which the method is applied to a computer device is configured for description. The computer device may be a terminal or a server. The method may be independently performed by the terminal or the server, or may be implemented through interaction between the terminal and the server.

[0041] Operation S202: Obtain a target video.

[0042] The target video is a to-be-analyzed video. The target video may be a pre-captured video. For example, the target video may be a security recording for demonstration. Alternatively, the target video may be a video captured in real time. For example, the target video may be a real-time caregiving video.

[0043] Specifically, the computer device may obtain the target video locally or from another device, analyze the target video, and recognize an object state of a target object in the target video.

[0044] Operation S204: Obtain a background frame corresponding to the target video, and determine, from the target video, a target frame corresponding to a target action of the target object.

[0045] The background frame is a background image, and is configured for representing a static or relatively constant background part in a scene. The background frame is mainly configured

for providing a sense of space and an ambient atmosphere in the scene, and helping recognize an environment in which the scene is located. The background frame corresponding to the target video is an image representing a background of a scene in the target video. The background frame corresponding to the target video may be a pre-captured background image. For example, the target video is a recording of a room A. Before video recording is started by using a camera, an image of the room A without any living object is captured by the same camera as the background frame corresponding to the target video. The background frame corresponding to the target video may be a background image determined from the target video. For example, a video frame without a living object is obtained from the target video as the background frame.

[0046] The target frame is an image representing that the target object performs the target action. In the target frame, the target object performs the target action. The target object is a to-be-analyzed object, and the target object may be specifically set according to a requirement. The target object is usually a living object. For example, the target object may be a person or an animal. The target action is a to-be-analyzed action, and the target action may be specifically set according to a requirement. For example, the target action may be a falling action or an eating action. For example, if the target object is a person, and the target action is a falling action, the person may present a lying posture in the target frame.

[0047] The background frame does not include the target object, and the target frame includes the target object.

[0048] Specifically, after determining the target video, the computer device may obtain the background frame corresponding to the target video, determine, from the target video, the target frame corresponding to the target action of the target object, and recognize, based on the background frame and the target frame, a target object state corresponding to the target object in the target frame.

[0049] For the target frame, the computer device may perform target object recognition on a video frame in the target video, further perform target action recognition on a video frame in which the target object is recognized, and use a video frame in which the target action is recognized as the target frame.

[0050] In an embodiment, target object recognition is performed on a video frame in the target video, and a video frame in which the target object is recognized is used as a candidate frame. Candidate frames within a target time period are obtained to form a candidate frame sequence. The candidate frame sequence is inputted into an action recognition model, to obtain a predicted action corresponding to the candidate frame sequence. When the predicted action is used as the target

action, the target frame corresponding to the target action of the target object is determined from the candidate frame sequence.

[0051] Operation S206: Perform image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame.

[0052] Image segmentation refers to dividing an image into several different regions. Pixels in the same region have a similar feature, and pixels in different regions have apparently different features.

[0053] Image segmentation is performed on the background frame to obtain the item region corresponding to each item in the background frame. A region corresponding to an item is a region in which the item is located in the image, and is configured for representing a position of the item. For example, if the background frame is an image of the room A, image segmentation may be performed on the background frame, to obtain a region in which each facility is located in the room A, to help understand distribution of facilities in the room A. Image segmentation is performed on the target frame, to obtain the target region corresponding to the target object in the target frame. The target region is a region in which the target object is located in the image, and is configured for representing a position of the target object.

[0054] The same image segmentation algorithm or different image segmentation algorithms may be used for the background frame and the target frame.

[0055] Specifically, the computer device may perform image segmentation on the background frame, to obtain the item region corresponding to each item in the background frame, and may perform image segmentation on the target frame, to obtain the target region corresponding to the target object in the target frame. For example, the background frame is inputted into an image segmentation model, and the image segmentation model outputs an image segmentation result of the background frame. The image segmentation result may be that an item is circled in the background frame using a positioning frame. The positioning frame may be a regular frame, or may be an irregular frame. An image region that is circled by using the positioning frame and that corresponds to an item is an item region corresponding to the item. For another example, the target frame is inputted into a target object recognition model, and the target object recognition model outputs the target region corresponding to the target object in the target frame. The target object recognition model concerns only a region in which the target object is located.

[0056] Operation S208: Generate an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item.

[0057] One item corresponds to one interaction information group. The interaction information group includes related information of the item in the background frame and related information of the target object in the target frame, and is configured for analyzing an interaction relationship between the target object and the item. The information about the item includes at least one of an item identifier corresponding to the item and the position information of the item region. The item identifier is configured for identifying an item. For example, the item identifier may be an item name or an item number. The position information of the item region is configured for indicating a position of the item region in the image. The related information of the target object includes at least one of an action identifier corresponding to the target action of the target object, the position information of the target region, and an object identifier corresponding to the target object. The action identifier is configured for identifying an action. For example, the action identifier may be an action name or an action number. The position information of the target region is configured for indicating a position of the target region in the image. The object identifier is configured for identifying an object. For example, the object identifier may be an object number.

[0058] Specifically, for any item in the background frame, the computer device may establish an interaction information group based on position information of an item region corresponding to the item, the item, the position information corresponding to the target region, and the target action. For example, the interaction information group includes a region overlap ratio between the item region and the target region, the item identifier of the item, and the action identifier of the target action; or the interaction information group includes the position information of the item region, the position information of the target region, the item identifier of the item, and the action identifier of the target action. Finally, the computer device may obtain the interaction information group between the target object in the target frame and each item in the background frame.

[0059] Operation S210: Determine, based on each interaction information group, a target object state corresponding to the target object in the target frame.

[0060] An object state is a state of an object because of performing an action. The same behavior may correspond to different object states in different situations. For example, a person may be in a rest state or a falling state when performing a falling action. When the falling action

highly overlaps furniture such as a bed, a sofa, and a recliner, the person is usually in the rest state, and when the falling action highly overlaps a floor, the person is usually in the falling state.

[0061] Specifically, the computer device may analyze the interaction relationship between the target object and the item based on each interaction information group, to determine the target object state corresponding to the target object in the target frame. For example, the target frame and the interaction information group are inputted into a trained interaction model, and the trained interaction model outputs the target object state corresponding to the target object in the target frame. The trained interaction model is obtained by performing supervised training on a to-be-trained interaction model based on a training frame, a training information group, and a training label corresponding to the training frame. For another example, a region overlap ratio between the target region and each item region is determined based on each interaction information group, an item corresponding to an item region whose region overlap ratio is greater than an overlap ratio threshold is obtained as a target item, and the target object state corresponding to the target object in the target frame is determined based on the target item and the target action.

[0062] In the foregoing object state recognition method, the target video is obtained; the background frame corresponding to the target video is obtained, and the target frame corresponding to the target action of the target object is determined from the target video; image segmentation is performed on the background frame and the target frame, to obtain the item region corresponding to each item in the background frame and the target region corresponding to the target object in the target frame; the interaction information group is generated based on the position information of the target region and the position information of the item region, the item corresponding to the item region, and the target action, to obtain the interaction information group between the target object and each item; and the target object state corresponding to the target object in the target frame is determined based on each interaction information group. In this way, the target action of the target object in the target frame is analyzed based on the background frame and the target frame, to recognize the target object state corresponding to the target object in the target frame. The background frame can provide information about an item, and the target frame can provide information about the target object. An interaction information group is established based on the information about the item and the information about the target object, and an interaction relationship between the target object and the item is analyzed by using the interaction information group, so that accuracy of object state recognition can be effectively improved. When an object state is recognized, an interaction relationship between the target object and an environment is considered, so that object states that are easily confused due to the same action can be effectively distinguished, thereby improving accuracy of object state recognition.

[0063] In an embodiment, the obtaining a background frame corresponding to the target video includes:

[0064] performing target object recognition on a video frame in the target video, and each time it is found that no target object is recognized in a target quantity of consecutive video frames, obtaining a last video frame from the target quantity of consecutive video frames as the background frame.

[0065] Target object recognition refers to recognizing whether a target object exists in an image. For example, if the target object is a person, human body recognition may be performed by using various human body detection algorithms. The target quantity may be set according to an actual requirement. The target quantity may be a preset quantity. Alternatively, the target quantity may be a quantity that flexibly changes with a scene.

[0066] Specifically, the computer device may determine the background frame from the target video, to improve accuracy of the background frame. The computer device may sequentially perform target object recognition on the video frame in the target video, to recognize whether the target object exists in the video frame. The computer device may count a quantity of consecutive video frames in which no target object is recognized. When no target object is detected in a target quantity of consecutive video frames, a last video frame is obtained from the target quantity of consecutive video frames in which no target object is detected as the background frame, to ensure that the background frame is a latest background image.

[0067] After determining the background frame, the computer device may clear a counting result, and restart counting a quantity of consecutive video frames in which no target object is recognized. When it is found again that no target object is detected in a target quantity of consecutive video frames, a last video frame is obtained from the target quantity of newly found consecutive video frames as a new background frame, to further ensure that the background frame is a latest background image all the time.

[0068] In a specific application, referring to FIG. 3, human body detection is performed on the target video frame by frame, to determine whether a human body is detected (that is, whether N_{object} is equal to 0). A quantity N_{empty} of consecutive frames in which no human body is detected is counted, and when no target object is detected in consecutive T_e frames, an image of a current frame f_i is updated as a background frame f_{bg} . When human body detection is performed, a target detection model such as a YOLO series target detector, a single shot multibox detector (SSD), CornerNet, or CenterNet may be used, but is not limited thereto.

[0069] In the foregoing embodiment, target object recognition is performed on the video frames in the target video, and each time it is found that no target object is recognized in the target quantity of consecutive video frames, the last video frame is obtained from the target quantity of consecutive video frames as the background frame. This can ensure that a latest and most accurate video frame representing a background in the target video is used as the background frame. Such a background frame helps improve accuracy of object state recognition.

[0070] In an embodiment, the target quantity is negatively correlated with recognition accuracy corresponding to a used target object recognition algorithm.

[0071] The target object recognition algorithm is an algorithm configured for performing target object recognition, and is an algorithm configured for recognizing a target object. The recognition accuracy corresponding to the target object recognition algorithm is recognition accuracy corresponding to the target object recognition algorithm.

[0072] Specifically, the target quantity may be a dynamically changed quantity, and may be specifically a quantity determined based on the target object recognition algorithm. The target quantity is related to the recognition accuracy of the target object recognition algorithm. Higher recognition accuracy of the target object recognition algorithm indicates a smaller target quantity. When no target object is recognized in the target quantity of consecutive video frames, the last frame is obtained as the background frame, so that a case in which the background frame includes the target object can be effectively avoided. If the recognition accuracy of the target object recognition algorithm configured for recognizing the target object is relatively poor, the target quantity can be increased, to improve accuracy of the background frame. When the recognition accuracy of the target object recognition algorithm is relatively poor, it is necessary to continuously recognize that more video frames do not have the target object, so as to determine the background frame.

[0073] In an embodiment, the target object recognition algorithm is determined based on a protection priority of the target object. A higher protection priority of the target object indicates that a target object recognition algorithm with higher recognition accuracy is used. Higher importance of the target object indicates a higher protection priority of the target object. For example, a protection priority of a person is higher than a protection priority of a cat or a dog. For another example, the older a person is, the higher protection priority of the person is. A protection priority of a patient is higher than a protection priority of an ordinary person.

[0074] In addition, object recognition may be first performed by using a general object recognition algorithm, a first recognized living object is used as the target object, then the target

object recognition algorithm is determined based on the protection priority of the target object, and target object recognition is performed on subsequent video frames based on the target object recognition algorithm, to improve accuracy of target object recognition. For example, it is first recognized, by using a living body recognition algorithm, whether a living body exists in a video frame. If it is first recognized that a human body exists in the video frame, it is accurately recognized, by using a human body recognition algorithm, whether a human body exists in a subsequent video frame, to improve accuracy of human body recognition.

[0075] In the foregoing embodiment, the target quantity is negatively correlated with the recognition accuracy corresponding to the used target object recognition algorithm. When the recognition accuracy of the target object recognition algorithm is relatively poor, it is necessary to continuously recognize that more video frames do not have the target object, so as to determine the background frame. This can ensure accuracy of the background frame.

[0076] In an embodiment, as shown in FIG. 4, the determining, from the target video, a target frame corresponding to a target action of the target object includes the following operations.

[0077] Operation S402: Perform target object recognition on a video frame in the target video, and use a video frame in which the target object is recognized as a candidate frame.

[0078] Specifically, the computer device may perform target object recognition on the video frame in the target video, use the video frame in which the target object is recognized as the candidate frame, further perform target action recognition on the candidate frame, and use a candidate frame in which the target action is recognized as the target frame.

[0079] In an embodiment, in a home monitoring or caregiving scenario, target object recognition is performed on a video frame in a target video, and a video frame in which a single target object is recognized is used as a candidate frame. In the home monitoring or caregiving scenario, the target object is usually a person. If there is more than one target object and one experiences an abnormal state, others on-site can provide assistance without the need for an automatic alarm.

[0080] Operation S404: Obtain candidate frames within a target time period to form a candidate frame sequence.

[0081] Execution of an action by the target object causes a posture change of the target object. Therefore, a corresponding action can be recognized by detecting the posture change of the target object. The posture change is time-related. Therefore, a plurality of video frames within a time period may be analyzed to recognize whether the target action occurs. For example, a falling

action represents a change from a standing posture to a lying posture of an object, and a sitting action represents a change from a standing posture to a sitting posture of an object. The target time period is a to-be-analyzed time period. The target time period may be set according to a requirement. A time length of the target time period is usually greater than or equal to implementation time of an action. For example, the target time period may be one second. It is determined, in candidate frames whose time spans are one second, whether the target object performs the target action. The candidate frame sequence includes a plurality of candidate frames whose time spans are within the target time period.

[0082] Specifically, the computer device may form the candidate frame sequence by using the candidate frames within the target time period in a chronological order, then perform target action recognition on the candidate frame sequence, and determine the target frame from the candidate frame sequence in which the target action exists.

[0083] Operation S406: Perform target object skeleton point recognition on the candidate frames in the candidate frame sequence, to obtain skeleton point data corresponding to the candidate frames, and form a skeleton point data sequence corresponding to the candidate frame sequence.

[0084] Target object skeleton point recognition refers to recognizing a skeleton point of the target object. For example, if the target object is a person, human body skeleton point recognition may be performed by using various human body skeleton point recognition algorithms. Target object skeleton point recognition is performed on a video frame, to obtain skeleton point data corresponding to the video frame. The skeleton point data sequence corresponding to the candidate frame sequence includes skeleton point data respectively corresponding to the candidate frames in the candidate frame sequence. The skeleton point data includes a type and a position of a skeleton point.

[0085] Specifically, for any candidate frame in the candidate frame sequence, the computer device may perform target object skeleton point recognition on the candidate frame, to obtain skeleton point data corresponding to the candidate frame. The computer device may obtain skeleton point data respectively corresponding to the candidate frames in the candidate frame sequence, and form the skeleton point data sequence corresponding to the candidate frame sequence by using the skeleton point data respectively corresponding to the candidate frames. For example, all pieces of skeleton point data may be directly combined to obtain the skeleton point data sequence. Alternatively, all pieces of skeleton point data belonging to the same type may be arranged in a chronological order, to obtain skeleton point time-sequence data corresponding to each type, and all

pieces of skeleton point time-sequence data are combined to obtain the skeleton point data sequence. Time corresponding to skeleton point data may be time corresponding to a candidate frame to which the skeleton point data belongs. The skeleton point data sequence can reflect an action feature of the candidate frame sequence, helping recognize an action corresponding to the candidate frame sequence.

[0086] Operation S408: Input the skeleton point data sequence into a multi-frame action recognition model, to obtain a predicted action corresponding to the candidate frame sequence.

[0087] The multi-frame action recognition model is an artificial intelligence model, and is configured to recognize an action based on a multi-frame image. Input data of the multi-frame action recognition model includes a skeleton point data sequence corresponding to an image sequence, and output data of the multi-frame action recognition model includes a predicted action. The predicted action is an action recognized by a model.

[0088] Specifically, the computer device may recognize the target action through the artificial intelligence model. The computer device obtains the multi-frame action recognition model, and inputs the skeleton point data sequence corresponding to the candidate frame sequence into the multi-frame action recognition model. The multi-frame action recognition model performs data processing on the inputted data, and outputs the predicted action corresponding to the candidate frame sequence.

[0089] The multi-frame action recognition model may be obtained by performing supervised training based on a skeleton point data sequence and an action training label that correspond to a training image sequence. The training image sequence is an image sequence of a known action. The action training label corresponding to the training image sequence is a known action reflected by the training image sequence.

[0090] Operation S410: When the predicted action is used as the target action, determine, from the candidate frame sequence, the target frame corresponding to the target action of the target object.

[0091] Specifically, when the predicted action outputted by the multi-frame action recognition model is used as the target action, the computer device may determine, from the candidate frame sequence, the target frame corresponding to the target action of the target object. For example, a last frame is obtained from the candidate frame sequence as the target frame. For another example, a frame is randomly selected from the candidate frame sequence as the target frame.

[0092] In the foregoing embodiment, target object recognition is performed on the video frame in the target video, the video frame in which the target object is recognized is used as the candidate frame, the candidate frames within the target time period are obtained to form the candidate frame sequence, target action recognition is performed on the candidate frame sequence, and the target frame is determined from the candidate frame sequence in which the target action is recognized. In this way, it is recognized, by comprehensively considering candidate frames within a time period, whether the target action occurs, so that accuracy of target action recognition can be effectively improved. Target object skeleton point recognition is performed on the candidate frames in the candidate frame sequence, to obtain the skeleton point data corresponding to the candidate frames, and the skeleton point data sequence corresponding to the candidate frame sequence is formed. The skeleton point data sequence is inputted into the multi-frame action recognition model, to obtain the predicted action corresponding to the candidate frame sequence. When the predicted action is used as the target action, the target frame corresponding to the target action of the target object is determined from the candidate frame sequence. In this way, the skeleton point data sequence corresponding to the candidate frame sequence is first extracted. The skeleton point data sequence can reflect an action feature. The action corresponding to the candidate frame sequence is recognized by processing the skeleton point data sequence through the multi-frame action recognition model. The multi-frame action recognition model has a powerful data processing capability, so that accuracy of target action recognition can be further improved.

[0093] In an embodiment, the determining, from the candidate frame sequence, the target frame corresponding to the target action of the target object includes:

[0094] separately inputting each candidate frame in the candidate frame sequence into a trained single-frame action recognition model, to obtain an action recognition result corresponding to each candidate frame, the action recognition result including a quantity and actions of target objects in the candidate frame; and obtaining a candidate frame whose action recognition result includes a single target object and the target action, and using the candidate frame as the target frame corresponding to the target action of the target object.

[0095] The single-frame action recognition model is an artificial intelligence model, and is configured to recognize an action based on a single-frame image. Input data of the single-frame action recognition model includes the single-frame image, and output data of the single-frame action recognition model includes a quantity and actions of target objects in the single-frame image.

[0096] Specifically, when performing target action recognition, the computer device may first perform pre-recognition through the multi-frame action recognition model, to filter out a video

frame apparently irrelevant to the target action. When the target action is pre-recognized, further recognition is performed through the single-frame action recognition model, to finally obtain the accurate target frame.

[0097] The computer device inputs the skeleton point data sequence corresponding to the candidate frame sequence into the multi-frame action recognition model, to obtain the predicted action corresponding to the candidate frame sequence. Affected by various factors such as computing resources, lighting, and angles, some false action recognition results may exist in initial screening results obtained through the multi-frame action recognition model, or even some data includes a plurality of target objects. Therefore, target action recognition is further performed through the single-frame action recognition model. When the predicted action corresponding to the candidate frame sequence is used as the target action, each candidate frame in the candidate frame sequence is separately inputted into the trained single-frame action recognition model, to obtain the action recognition result corresponding to each candidate frame. The action recognition result corresponding to the candidate frame includes the quantity and actions of target objects in the candidate frame. For example, the action recognition result may be an action of circling the target object by using a positioning frame and marking the target object in the candidate frame. The computer device obtains, from the candidate frame sequence, the candidate frame whose action recognition result includes the single target object and the target action, and uses the candidate frame as the target frame corresponding to the target action of the target object.

[0098] In the foregoing embodiment, the single-frame action recognition model can not only recognize the quantity of target objects, but also recognize the target action. The single-frame action recognition model implements functions of filtering a multi-object scenario and avoiding false action recognition, so that recognition accuracy of an effective target frame can be improved.

[0099] In a specific application, referring to FIG. 5, video frames whose time spans are within 1s are obtained, the video frames are arranged in a chronological order and are inputted into a multi-frame action recognition model, and if a recognition result outputted by the multi-frame action recognition model is a target action, the video frames are separately inputted into a single-frame action recognition model. If the recognition result outputted by the multi-frame action recognition model is not the target action, it may be considered that there is no target frame in these video frames. The single-frame action recognition model can recognize a quantity and actions of target objects in a video frame. If a recognition result outputted by the single-frame action recognition model for a video frame is a single target object and a target action, the video frame is used as a target frame.

[00100] In an embodiment, the object state recognition method further includes:

[00101] obtaining a first training image, and a target object position training label and a target object action training label that correspond to the first training image; inputting the first training image into a to-be-trained single-frame action recognition model, to obtain a target object position prediction label and a target object action prediction label that correspond to the first training image; and adjusting a model parameter of the to-be-trained single-frame action recognition model based on a difference between the target object position training label and the target object position prediction label and a difference between the target object action training label and the target object action prediction label, until a first convergence condition is met, to obtain the trained single-frame action recognition model.

[00102] The first training image is an image configured for training the single-frame action recognition model. The target object position training label corresponding to the first training image is configured for indicating accurate position information of the target object in the first training image. The target object action training label corresponding to the first training image is configured for indicating accurate action information of the target object in the first training image. There may be no target object, only a single target object, or a plurality of target objects in the first training image. For example, if the target object is a person, there may be no person, only one person, or a plurality of persons in the first training image. The plurality of persons in the first training image may be persons of the same type or may be persons of different types. Alternatively, if the target object is a dog, there may be no dog, only one dog, or a plurality of dogs in the first training image. One target object in the first training image has a corresponding target object action training label. The target object action training label may be a target action, or may be another action. For example, the target object action training label may be a sitting posture, a standing posture, or a lying posture.

[00103] The first training image is inputted into the to-be-trained single-frame action recognition model, to obtain the target object position prediction label and the target object action prediction label that correspond to the first training image. The target object position prediction label is configured for indicating predicted position information of the target object in the first training image. The target object action prediction label is configured for indicating predicted action information of the target object in the first training image.

[00104] The first convergence condition is a condition for determining whether the single-frame action recognition model reaches convergence. The first convergence condition may be set according to an actual requirement. The first convergence condition may be at least one of the following: a model loss is less than a loss threshold, a quantity of model iterations is greater than an

iteration quantity threshold, a model loss change rate is less than a loss change rate threshold, and the like.

[00105] Specifically, a computer device may obtain the first training image, and the target object position training label and the target object action training label that correspond to the first training image, and perform supervised training on the single-frame action recognition model based on the first training image and the training labels corresponding to the first training image.

[00106] The computer device may input the first training image into the to-be-trained single-frame action recognition model. After performing data processing, the single-frame action recognition model outputs the target object position prediction label and the target object action prediction label that correspond to the first training image. Further, the computer device calculates a model loss based on the difference between the target object position training label and the target object position prediction label and the difference between the target object action training label and the target object action prediction label, and performs backpropagation on the model loss to adjust the model parameter of the to-be-trained single-frame action recognition model until the first convergence condition is met, to obtain the trained single-frame action recognition model.

[00107] There may be a plurality of first training images. In each round, a plurality of first training images may be obtained to participate in training. In each round, a model sub-loss is calculated based on a difference between a target object position training label and a target object position prediction label that correspond to the same first training image and a difference between a target object action training label and a target object action prediction label that correspond to the same first training image, to obtain a model sub-loss corresponding to each first training image. The model loss is obtained based on the model sub-loss corresponding to each training image. The to-be-trained single-frame action recognition model is adjusted based on the model loss, and an adjusted single-frame action recognition model is used as a to-be-trained single-frame action recognition model in a next round.

[00108] In the foregoing embodiment, supervised training is performed on the single-frame action recognition model based on the first training image, and the target object position training label and the target object action training label that correspond to the first training image, so that the trained single-frame action recognition model has a target object recognition capability and an action recognition capability.

[00109] In a specific application, referring to FIG. 6, a process of determining a target object state is described by using an example in which a target object is a human body and a target action is a falling action. A background frame and a target frame are determined from a target video. Image

semantic segmentation is separately performed on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a human body region corresponding to a human body in the target frame. An interaction information group is established based on the item region corresponding to the item and the human body region. One interaction information group includes an action identifier (namely, P_{falling} in FIG. 6) of the falling action, an item identifier (namely, a floor O_1 or a sofa O_2 in FIG. 6) of an item in the background frame, and an overlapping area (namely, E_1 or E_2 in FIG. 6) between the item region and the human body region. A target object state corresponding to the target object in the target frame is determined based on each interaction information group.

[00110] Image semantic segmentation may be performed through a deep neural network model, including but not limited to a Light-Weight RefineNet or a segment anything model (SAM).

[00111] In an embodiment, as shown in FIG. 7, the determining, based on each interaction information group, a target object state corresponding to the target object in the target frame includes the following operations.

[00112] Operation S702: Obtain reference object states corresponding to the target object under the target action, and obtain reference items respectively corresponding to the reference object states.

[00113] The target object may have at least one reference object state under the target action. The reference object state is an object state of the target object under the target action. For example, for a falling action, a reference object state corresponding to the target object under the falling action may be a rest state or may be a falling state; for a sitting action, a reference object state corresponding to the target object under the sitting action may be a rest state or may be a social state; and for a running action, a reference object state corresponding to the target object under the running action may be an emergency-avoidance state or may be an exercising state.

[00114] The reference object state has a corresponding reference item. The reference item is an item related to the reference object state. For example, when the target object is in the rest state, the reference item may be an item such as a bed, a sofa, or a chair; when the target object is in the falling state, the reference item may be an item such as a floor or a step; and when the target object is in the exercising state, the reference item may be an item such as a treadmill or a sandbag.

[00115] Specifically, the computer device may obtain the reference object states corresponding to the target object under the target action, and determine the target object state corresponding to the target object in the target frame from the reference object states.

[00116] Operation S704: Use a reference item in the background frame as a target item, and obtain a reference overlap ratio threshold corresponding to a reference object state corresponding to the target item as a target overlap ratio threshold.

[00117] The reference object state has a corresponding reference overlap ratio threshold. The reference overlap ratio threshold is a condition that needs to be met by an overlap ratio between the target object in the reference object state and the reference item. If an overlap ratio between the target object under the target action and the reference item is greater than the reference overlap ratio threshold, it is determined that the target object under the target action is in the reference object state corresponding to the reference item. For example, for the falling action, if an overlap ratio between the target object and an item such as a bed, a sofa, or a chair is greater than a threshold, it is determined that the target object is in the rest state; or if an overlap ratio between the target object and an item such as a floor or a step is greater than a threshold, it is determined that the target object is in the falling state. For the running action, if an overlap ratio between the target object and an item such as a treadmill or a runway is greater than a threshold, it is determined that the target object is in the exercising state. The reference overlap ratio threshold corresponding to the reference object state may be obtained by performing statistical analysis on a large quantity of known images in which target objects are in the reference object state. For example, a region overlap ratio between a target region corresponding to a target object in a known image and an item region corresponding to a reference item is calculated, and an average value of region overlap ratios respectively corresponding to the known images is calculated and used as the reference overlap ratio threshold, or a median value of region overlap ratios respectively corresponding to the known images is obtained and used as the reference overlap ratio threshold.

[00118] The target item is the reference item included in the background frame. For example, if the target action is the falling action, the target item may be an item such as a bed, a sofa, a chair, a floor, or a step in the background frame. The target overlap ratio threshold is a reference overlap ratio threshold corresponding to the reference object state corresponding to the target item.

[00119] Specifically, the computer device may obtain the reference items respectively corresponding to the reference object states, use a reference item that exists in the background frame as the target item, and determine, from the reference object state corresponding to the target item, the target object state corresponding to the target object in the target frame.

[00120] Operation S706: For an interaction information group corresponding to the target item, when a region overlap ratio between an item region corresponding to the target item and the target region is greater than the target overlap ratio threshold, use the reference object state

corresponding to the target item as the target object state corresponding to the target object in the target frame.

[00121] The region overlap ratio between the item region and the target region is a degree of overlapping between the item region and the target region and may be specifically an overlapping area, an overlapping proportion, or the like.

[00122] Specifically, the computer device obtains the reference overlap ratio threshold corresponding to the reference object state corresponding to the target item as the target overlap ratio threshold, determines, based on the interaction information group corresponding to the target item, the region overlap ratio between the item region corresponding to the target item and the target region corresponding to the target object, and uses the reference object state corresponding to the target item as the target object state corresponding to the target object in the target frame if the region overlap ratio is greater than the target overlap ratio threshold.

[00123] If there are a plurality of reference items for one reference object state, and a region overlap ratio between the target region and an item region corresponding to any one of the reference items is greater than a corresponding reference overlap ratio threshold, the reference object state corresponding to the reference item is used as the target object state.

[00124] If the region overlap ratio is less than or equal to the target overlap ratio threshold, the target object state corresponding to the target object in the target frame may be further determined in another manner. For example, the target object state corresponding to the target object in the target frame is determined through a trained interaction model. The target frame and each interaction information group are inputted into the trained interaction model, and the trained interaction model outputs the target object state corresponding to the target object in the target frame. In this way, it is first determined, based on related information of the reference object state, whether the reference object state is hit, so that efficiency of object state recognition can be improved. If the reference object state is not hit, an object state is recognized through the interaction model, so that accuracy of object state recognition can be improved, and shortcomings of the reference object state can be compensated. For another example, the target frame is transmitted to a management terminal, and the management terminal displays the target frame to an administrator, so that the administrator determines, based on the target frame, the target object state corresponding to the target object in the target frame. If the region overlap ratio is less than or equal to the target overlap ratio threshold, the target object state corresponding to the target object in the target frame may alternatively be determined as an unknown state.

[00125] In the foregoing embodiment, the reference object states corresponding to the target object under the target action are obtained. The reference object state has a corresponding reference item and a reference overlap ratio threshold. The region overlap ratio between the item region corresponding to the reference item included in the background frame and the target region corresponding to the target object in the target frame is determined, and the region overlap ratio is compared with the corresponding reference overlap ratio threshold, so that the target object state corresponding to the target object in the target frame can be quickly determined. In addition, each reference object state has a corresponding reference overlap ratio threshold, which helps further improve accuracy of object state recognition.

[00126] In an embodiment, as shown in FIG. 8, the determining, based on each interaction information group, a target object state corresponding to the target object in the target frame includes the following operations.

[00127] Operation S802: Input the target frame and each interaction information group into a trained interaction model.

[00128] The interaction model is an artificial intelligence model, and is configured to determine an object state based on an image and the interaction information group. Input data of the interaction model includes an image corresponding to an action of an object, and an interaction information group determined based on the image and a corresponding background image. Output data includes an object state recognized by the model.

[00129] Specifically, the computer device may recognize the target object state through the artificial intelligence model. The computer device obtains the trained interaction model, and inputs the target frame corresponding to the target action of the target object and each interaction information group determined based on the background frame and the target frame into the trained interaction model. After performing data processing, the interaction model outputs the target object state corresponding to the target object in the target frame.

[00130] Operation S804: Separately determine an overlapping region in the target frame based on each interaction information group.

[00131] Operation S806: Extract an image feature of the overlapping region, to obtain a first interaction feature, obtain a second interaction feature based on an item and a target action that are in an interaction information group corresponding to the overlapping region, and extract an image feature of the target frame, to obtain a third interaction feature.

[00132] Operation S808: Obtain a comprehensive interaction feature based on the first interaction feature, the second interaction feature, and the third interaction feature.

[00133] Operation S810: Output, based on the comprehensive interaction feature, the target object state corresponding to the target object in the target frame.

[00134] The overlapping region is an overlapping part between an item region corresponding to an item in the background frame and the target region corresponding to the target object in the target frame. The first interaction feature is the image feature of the overlapping region, and includes information about the overlapping region. The second interaction feature is a feature determined based on the item and the target action that are in the interaction information group corresponding to the overlapping region, and includes information about the item and the target action that correspond to the overlapping region. The third interaction feature is the image feature of the target frame, and includes information about the target frame. The comprehensive interaction feature is a feature obtained based on the first interaction feature, the second interaction feature, and the third interaction feature, and includes important information in the three types of interaction features.

[00135] Specifically, the computer device inputs the target frame and each interaction information group into the trained interaction model. In the interaction model, the overlapping region is separately determined in the target frame based on each interaction information group, to obtain at least one overlapping region. The image feature of the overlapping region is extracted to obtain the first interaction feature, and the image feature of the target frame is extracted to obtain the third interaction feature. For example, an image feature may be extracted through a convolution layer. In the interaction model, the second interaction feature is generated based on the item and the target action that are in the interaction information group corresponding to the overlapping region. For example, a feature of an item is extracted, a feature of the target action is extracted, and the extracted features are concatenated, to obtain the second interaction feature. In the interaction model, the comprehensive interaction feature is obtained based on the first interaction feature, the second interaction feature, and the third interaction feature. For example, the first interaction feature, the second interaction feature, and the third interaction feature are fused to obtain the comprehensive interaction feature. Feature fusion may be performed through a fully connected layer. Alternatively, attention processing is separately performed on every two of the first interaction feature, the second interaction feature, and the third interaction feature, to obtain a plurality of attention features, and the plurality of attention features are fused to obtain the comprehensive interaction feature. Attention processing may be performed by using an attention mechanism, and

some features that affect each other may be extracted through attention processing. In the interaction model, the target object state corresponding to the target object in the target frame is determined based on the comprehensive interaction feature. The interaction model finally outputs the target object state corresponding to the target object in the target frame.

[00136] In the foregoing embodiment, the target frame and each interaction information group are inputted into the trained interaction model. The interaction model has a powerful data processing capability and can quickly output the accurate target object state. In addition, the interaction information group can provide interaction information between the target object and the item, and help the interaction model understand the target frame. Therefore, the target object state outputted by the interaction model further has higher accuracy.

[00137] In an embodiment, the object state recognition method further includes:

[00138] obtaining a second training image corresponding to a training action of a training object, and a training object state and a training information group that correspond to the second training image, the training object state being an object state corresponding to the training object in the second training image, and the training information group being determined based on position information of a region in which the training object is located in the second training image, position information of a region in which an item is located in a background image corresponding to the second training image and item information, and the training action; inputting the second training image and the training information group into a to-be-trained interaction model, to obtain a predicted object state corresponding to the second training image; and adjusting a model parameter of the to-be-trained interaction model based on a difference between the training object state and the predicted object state, until a second convergence condition is met, to obtain the trained interaction model.

[00139] The second training image is an image configured for training the interaction model. The second training image is an image corresponding to the training action of the training object. The training object state corresponding to the second training image is configured for indicating an accurate object state corresponding to the training object in the training image. The training information group is an interaction information group determined based on an image segmentation result of the second training image and an image segmentation result of the background image corresponding to the second training image. For a method for determining the training information group, refer to the foregoing method for determining the interaction information group.

[00140] The second training image and the training information group corresponding to the second training image are inputted into the to-be-trained interaction model, to obtain the predicted

object state corresponding to the second training image. The predicted object state is an object state that corresponds to the training object in the second training image and that is predicted by the model.

[00141] The second convergence condition is a condition for determining whether the interaction model reaches convergence. The second convergence condition may be set according to an actual requirement. The second convergence condition may be at least one of the following: a model loss is less than a loss threshold, a quantity of model iterations is greater than an iteration quantity threshold, a model loss change rate is less than a loss change rate threshold, and the like.

[00142] The training object and the target object may be the same object, or may be different objects. The training action and the target action may be the same action, or may be different actions. The first training image and the second training image may be the same training image, or may be different training images.

[00143] Specifically, the computer device may obtain the second training image corresponding to the training action of the training object, and the training object state and the training information group that correspond to the second training image, and perform supervised training on the interaction model based on the data. The computer device inputs the second training image and the training information group into the to-be-trained interaction model. After performing data processing, the interaction model outputs the predicted object state corresponding to the second training image. A model loss is further calculated based on a difference between the training object state and the predicted object state, and the model loss is backpropagated to adjust the model parameter of the to-be-trained interaction model until the second convergence condition is met, to obtain the trained interaction model.

[00144] There may be a plurality of second training images. In each round, a plurality of second training images may be obtained to participate in training.

[00145] In the foregoing embodiment, supervised training is performed on the interaction model based on the second training image, and the training object state and the training information group that correspond to the second training image, so that the trained interaction model has an object state recognition capability.

[00146] In an embodiment, the target video is a video stream obtained by capturing an image of a living environment of the target object in real time. The object state recognition method further includes:

[00147] when the target object state is an abnormal object state, transmitting an alarm message for the abnormal object state to a management terminal corresponding to the living environment.

[00148] The target video is the video stream obtained by capturing the image of the living environment of the target object in real time. For example, the target video may be a monitoring video captured in real time by using a monitoring camera installed in a home of the target object; the target video may be a monitoring video captured in real time by using a monitoring camera installed in an elderly care institution or a hospital in which the target object is located.; or the target video may be a monitoring video captured in real time by using a monitoring camera installed in a place frequently visited by the target object.

[00149] The abnormal object state is an abnormal object state. The abnormal object state may be set according to a requirement. For example, a dangerous object state may be used as the abnormal object state; or an object state different from a daily object state may be used as the abnormal object state. For example, life of the target object is regular, and an object state of the target object within each time period of a day is basically stable and regular. If a current object state of the target object within a time period is different from a historical object state within the time period, it may be determined that the current object state of the target object within the time period is the abnormal object state.

[00150] The management terminal corresponding to the living environment is a management terminal for the target object, and is configured to manage an object state of the target object in the living environment. For example, if the living environment of the target object is a home environment of the target object, the management terminal may be a family, a guardian, or the like of the target object. If the living environment of the target object is a living environment of the target object in an elderly care institution or a hospital, the management terminal may be a staff member of the elderly care institution or the hospital, or may be a family, a guardian, or the like of the target object.

[00151] The alarm message for the abnormal object state is configured for alarming that the target object is in the abnormal object state. The alarm message may be a message in at least one media form such as a text, an image, a video, or audio. The alarm message may be displayed in various manners such as an SMS message, an application program, and a call.

[00152] Specifically, the target video is the video stream obtained by capturing the image of the living environment of the target object in real time. In other words, the target video is a continuously generated video stream. The computer device may obtain the target video and decode

the target video to obtain a video frame for object state recognition. If there is the target frame corresponding to the target action of the target object, the target object state corresponding to the target object in the target frame is determined based on the target frame and the background frame, and it is further recognized whether the target object state is the abnormal object state. If the target object state is the abnormal object state, the computer device may transmit the alarm message for the abnormal object state to the management terminal corresponding to the living environment, so that the management terminal can learn in time that the target object is in the abnormal object state, and take corresponding protection measures in time.

[00153] In the foregoing embodiment, when the target object state is the abnormal object state, the alarm message for the abnormal object state is transmitted to the management terminal corresponding to the living environment, thereby helping the target object return to a normal object state in time, and ensuring security of the target object.

[00154] In an embodiment, the target video is a video stream obtained by capturing, by a target device arranged in a living environment of the target object, an image of the living environment in real time. The object state recognition method further includes:

[00155] when the target object state is an abnormal object state, transmitting an alarm message for the abnormal object state to the target device, the target device being configured to: display an assistance requirement confirmation message for the abnormal object state, determine a target assistance manner for the target object in response to an assistance acceptance consent event for the assistance requirement confirmation message, and transmit an assistance message for the target object to an assistance device corresponding to the target assistance manner.

[00156] The assistance requirement confirmation message is a message configured for determining whether the target object needs assistance. The assistance requirement confirmation message may also be referred to as an assistance requirement query message. If the target object needs assistance, the target object interacts with the target device, so that the target device transmits an assistance message for the target object to another device. For example, the assistance requirement confirmation message may be a voice message querying whether the target object needs assistance. The target object may dialogue with the target device to feed back whether the target object needs assistance. For another example, the assistance requirement confirmation message may be a notification message querying whether the target object needs assistance. The target object may perform an operation on the target device to feed back whether the target object needs assistance.

[00157] The assistance acceptance consent event is an event representing that assistance is agreed to be accepted or that a need for assistance is confirmed. The assistance acceptance consent event may be generated based on a trigger operation of a user. For example, in response to a trigger operation for the assistance requirement confirmation message, the assistance acceptance consent event is triggered; or an agreement control is displayed, and in response to a trigger operation for the agreement control, the assistance acceptance consent event is triggered. The assistance acceptance consent event may alternatively be generated based on a mid-air operation of the user. The mid-air operation is an operation of transferring information to the target device in a non-contact manner. For example, the assistance requirement confirmation message may be a voice message querying whether the target object needs assistance. The target object may transmit a voice message representing that the target object agrees to receive assistance. If the target device receives the voice message transmitted by the target object, the target device generates the assistance acceptance consent event. Certainly, the assistance acceptance consent event may alternatively be automatically triggered. For example, if a user does not refuse to accept assistance within a time period, the assistance acceptance consent event is automatically triggered.

[00158] The target assistance manner for the target object is a specific manner of assisting the target object when the target object is in the abnormal object state. The target assistance manner may be a preset assistance manner. For example, the target assistance manner includes any one of an emergency contact assistance manner, a professional institution assistance manner, or a robot assistance manner. For another example, the target assistance manner may be an assistance manner selected from candidate assistance manners. The assistance requirement confirmation message may further include a plurality of candidate assistance manners for selection by the target object. If the target object needs assistance, the target object may select a candidate assistance manner, and the candidate assistance manner selected by the target object is used as the target assistance manner. The target object may select the target assistance manner from the candidate assistance manners in a contact or non-contact manner. Certainly, the target assistance manner may alternatively be a user-defined assistance manner. The target object may provide the user-defined assistance manner for the target device. For example, the target object may provide contact information of a friend for the target device, so that the target device transmits an assistance message for the target object to a friend terminal of the friend based on the contact information.

[00159] The target device transmits the assistance message for the target object to the assistance device corresponding to the target assistance manner. The assistance message for the target object includes position information of the target object and the target object state. For example, the candidate assistance manner includes the emergency contact assistance manner, the

professional institution assistance manner, and the robot assistance manner. If the target object selects the emergency contact assistance manner, the assistance device may be a communication device corresponding to an emergency contact of the target object. If the target object selects the professional institution assistance manner, the assistance device may be a communication device corresponding to a professional institution. If the target object selects the robot assistance manner, the assistance device is a robot whose distance from the living environment is within a range. The assistance message for the target object is transmitted to the robot, so that the robot can find the target object based on the position information of the target object, and take corresponding measures based on the target object state of the target object to help the target object.

[00160] In this way, when it is detected that the target object is abnormal, promptly querying, by using the assistance requirement confirmation message, whether the target object needs assistance can effectively help the target object return to normal as soon as possible.

[00161] In an embodiment, the target video is a video stream obtained by capturing an image of a living environment of the target object in real time. The object state recognition method further includes:

[00162] obtaining an adjacent frame set of the target frame from the target video when the target object state is an abnormal object state; performing target object recognition on an adjacent frame in the adjacent frame set, and using an adjacent frame in which the target object is recognized as an auxiliary frame; determining an auxiliary object from the auxiliary frame, the auxiliary object and the target object in the target frame being different objects of the same type; and transmitting an alarm message for the abnormal object state to an object terminal corresponding to the auxiliary object.

[00163] The adjacent frame set of the target frame is a video frame whose capture time distance from the target frame is less than a time distance threshold in the target video. In other words, the adjacent frame set of the target frame includes several video frames preceding and following the target frame in the target video.

[00164] The auxiliary frame is a video frame selected from the adjacent frame set of the target frame. The auxiliary object is another target object in the auxiliary frame that is different from the target object in the abnormal object state. For example, if the target object is a person, and Zhang San in the target frame is in the abnormal object state, the auxiliary object may be Li Si in the auxiliary frame. When it is detected that the target object in the target frame is in the abnormal object state, it is not detected that the auxiliary object is in the abnormal object state.

[00165] The object terminal corresponding to the auxiliary object is a communication device corresponding to the auxiliary object. For example, the object terminal corresponding to the auxiliary object may be a mobile phone or a smartwatch of the auxiliary object.

[00166] Specifically, when the target object state corresponding to target object in the target frame is the abnormal object state, the computer device may further determine the auxiliary object from the target video, and transmit the alarm message for the abnormal object state to the object terminal corresponding to the auxiliary object, so that the auxiliary object takes measures in time to help the target object.

[00167] When the target object state is the abnormal object state, the adjacent frame set of the target frame is obtained from the target video. For example, video frames within 5 minutes preceding and following the target frame are obtained to form the adjacent frame set. Further, target object recognition is performed on the adjacent frame in the adjacent frame set, the adjacent frame in which the target object is recognized is used as the auxiliary frame, and the auxiliary object is determined from the auxiliary frame. For example, an adjacent frame that is closest to the target frame and in which the target object is recognized is used as the auxiliary frame, and at least one other target object different from the target object in the target frame that is in the abnormal object state is selected from the auxiliary frame as the auxiliary object. Further, the alarm message for the abnormal object state is transmitted to the object terminal corresponding to the auxiliary object.

[00168] For example, the target object is a person. When a user A in the target frame is in the abnormal object state, if a user B appears in the adjacent frame set of the target frame, the user B is used as the auxiliary object. An alarm message for the user A is transmitted to a communication device corresponding to the user B, so that the user B learns in time that the user A has an abnormality through the communication device.

[00169] In the foregoing embodiment, when the target object state is the abnormal object state, the auxiliary object is determined from the adjacent frame set of the target frame, and the alarm message for the abnormal object state is transmitted to the object terminal corresponding to the auxiliary object. The auxiliary object is usually close to the target object, thereby helping quickly help the target object return to a normal object state, and ensuring security of the target object.

[00170] In a specific embodiment, the method of this application may be applied to falling detection in a home scenario, thereby effectively improving falling recognition accuracy. Referring to FIG. 9 and FIG. 10, a server obtains a video stream captured by using a camera in a home scenario and decodes the video stream into images. Human shape detection (namely, human body

detection) is first performed on the images. Once a target person is detected, an action performed by the target person in an image frame is recognized by using an action recognition module. In addition, a background frame in the video stream may also be recognized (namely, environmental context detection) based on a human body detection result of each image. An image in which it is recognized that the target person performs a falling action is used as a foreground frame (namely, a target frame). Image semantic segmentation is separately performed on the foreground frame and the background frame. In the foreground frame, attention is mainly paid to a pixel-level semantic segmentation region of the target person, and in the background frame, attention is paid to a segmentation result of an item overlapping or connected to a region of the target person. An overlap ratio between the target person and each item in a surrounding environment is calculated based on an image semantic segmentation result. The overlap ratio, the falling action, and the item together constitute an interaction triplet (namely, an interaction information group). A falling recognition result of the target person is determined based on the interaction triplet. The interaction triplet may be inputted into an interaction model, and the interaction model outputs a final falling recognition result.

[00171] If the falling recognition result is that the target person falls, the server may return the falling recognition result to a front end, and the front end may perform further operations such as image display and abnormal behavior alarm based on the falling recognition result. For example, the front end may display the foreground frame, and the target person who falls is circled in the foreground frame.

[00172] The target person may be set according to a requirement. For example, the target person may be an elderly person or a child.

[00173] With the growing elderly population, more advanced household devices are needed to monitor activities of elderly people, to ensure safety of the elderly people. According to the method of this application, in a region such as a bedroom or a living room, when there is furniture such as a bed and a sofa that are strongly associated with the falling action, falling actions of the elderly person in different situations can be more effectively distinguished, and whether the elderly person is in a rest state or a falling state can be distinguished.

[00174] Certainly, the method of this application is not only applied to home security protection, but also applied to scenarios such as patient assistance and care in a medical/elderly care institution. The method of this application may also be configured for detecting other states in addition to falling.

[00175] In a specific embodiment, the method of this application may be applied to a cloud-edge collaboration system, and falling detection is performed by using the cloud-edge collaboration system. The cloud-edge collaboration system integrates cloud computing and edge computing, aiming to implement efficient data processing and transmission. The cloud-edge collaboration system includes a cloud device and an edge device. The edge device, namely, an intelligent camera device, is configured to capture a real-time video and decode the real-time video into images, and complete preliminary analysis, including human body detection, human body skeleton point detection, action recognition based on skeleton point time sequence data, and the like, in a local device, to obtain an interaction triplet, and determine, based on the interaction triplet, a preliminary determining result about whether a target person falls. For example, if an overlap ratio in the interaction triplet is greater than a corresponding threshold, it is determined that the preliminary determining result is that the target person falls. When the preliminary determining result is that the target person falls, a result obtained through local processing, a current frame (namely, a foreground frame), and the interaction triplet are uploaded to the cloud device. The cloud device performs secondary analysis on the data transmitted by the edge device, to obtain a final determining result. For example, the foreground frame and the interaction triplet are inputted into an interaction model, and the interaction model outputs the final determining result. When the final determining result is that the target person falls, the cloud device can push the final determining result (for example, a falling alarm and a related image) to a user terminal or a monitoring center based on service logic. The cloud device can also automatically call an emergency contact person number of the target person.

[00176] In this way, efficient and accurate falling detection is implemented through collaboration of the edge device and the cloud device. The edge device performs preliminary determining to implement data filtering. Only when a preliminary determining result is that the target person falls, the edge device transmits related data to the cloud device, and the cloud device performs secondary determining to obtain a final determining result.

[00177] Although the operations in the flowcharts involved in the foregoing embodiments are sequentially shown according to indication of an arrow, the operations are not necessarily sequentially performed according to a sequence indicated by the arrow. Unless explicitly specified in this application, execution of the operations is not strictly limited, and the operations may be performed in another sequence. Moreover, at least a part of operations in the flowcharts involved in the foregoing embodiments may include a plurality of operations or a plurality of stages. These operations or stages are not necessarily executed at the same moment, but can be executed at different moments. The order of execution of these operations or stages is not necessarily performed

sequentially, but may be performed in turn or alternately with other operations or at least a part of operations or stages of other operations.

[00178] Based on the same inventive concept, an embodiment of this application further provides an object state recognition apparatus configured to implement the foregoing object state recognition method. An implementation solution provided by the apparatus for resolving a problem is similar to the implementation solution recorded in the foregoing method. Therefore, for specific limitations of one or more embodiments of the object state recognition apparatus provided below, refer to the limitations of the foregoing object state recognition method, and details are not described herein again.

[00179] In an embodiment, as shown in FIG. 11, an object state recognition apparatus is provided. The apparatus includes: a video obtaining module 1102, an image determining module 1104, an image segmentation module 1106, an information group generation module 1108, and a state recognition module 1110.

[00180] The video obtaining module 1102 is configured to obtain a target video.

[00181] The image determining module 1104 is configured to: obtain a background frame corresponding to the target video, and determine, from the target video, a target frame corresponding to a target action of a target object.

[00182] The image segmentation module 1106 is configured to perform image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame.

[00183] The information group generation module 1108 is configured to generate an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item.

[00184] The state recognition module 1110 is configured to determine, based on each interaction information group, a target object state corresponding to the target object in the target frame.

[00185] In an embodiment, the image determining module 1104 is further configured to:

[00186] perform target object recognition on a video frame in the target video, and each time it is found that no target object is recognized in a target quantity of consecutive video frames, obtain a last video frame from the target quantity of consecutive video frames as the background frame.

- [00187] In an embodiment, the image determining module 1104 is further configured to:
- [00188] perform object recognition on the video frame in the target video based on a general object recognition algorithm, and use a first recognized living object as the target object;
- [00189] determine a target object recognition algorithm based on a protection priority of the target object; and
- [00190] perform target object recognition on the video frame in the target video based on the target object recognition algorithm.
- [00191] In an embodiment, the target quantity is negatively correlated with recognition accuracy corresponding to the used target object recognition algorithm.
- [00192] In an embodiment, the image determining module 1104 is further configured to:
- [00193] perform target object recognition on the video frame in the target video, and use a video frame in which the target object is recognized as a candidate frame;
- [00194] obtain candidate frames within a target time period to form a candidate frame sequence;
- [00195] perform target object skeleton point recognition on the candidate frames in the candidate frame sequence, to obtain skeleton point data corresponding to the candidate frames, and form a skeleton point data sequence corresponding to the candidate frame sequence;
- [00196] input the skeleton point data sequence into a multi-frame action recognition model, to obtain a predicted action corresponding to the candidate frame sequence; and
- [00197] when the predicted action is used as the target action, determine, from the candidate frame sequence, the target frame corresponding to the target action of the target object.
- [00198] In an embodiment, the image determining module 1104 is further configured to:
- [00199] separately input each candidate frame in the candidate frame sequence into a trained single-frame action recognition model, to obtain an action recognition result corresponding to each candidate frame, the action recognition result including a quantity and actions of target objects in the candidate frame; and
- [00200] obtain a candidate frame whose action recognition result includes a single target object and the target action, and use the candidate frame as the target frame corresponding to the target action of the target object.
- [00201] In an embodiment, the object state recognition apparatus is further configured to:

[00202] obtain a first training image, and a target object position training label and a target object action training label that correspond to the first training image;

[00203] input the first training image into a to-be-trained single-frame action recognition model, to obtain a target object position prediction label and a target object action prediction label that correspond to the first training image; and

[00204] adjust a model parameter of the to-be-trained single-frame action recognition model based on a difference between the target object position training label and the target object position prediction label and a difference between the target object action training label and the target object action prediction label, until a first convergence condition is met, to obtain the trained single-frame action recognition model.

[00205] In an embodiment, the state recognition module 1110 is further configured to:

[00206] obtain reference object states corresponding to the target object under the target action, and obtain reference items respectively corresponding to the reference object states;

[00207] use a reference item in the background frame as a target item, and obtain a reference overlap ratio threshold corresponding to a reference object state corresponding to the target item as a target overlap ratio threshold; and

[00208] for an interaction information group corresponding to the target item, when a region overlap ratio between an item region corresponding to the target item and the target region is greater than the target overlap ratio threshold, use the reference object state corresponding to the target item as the target object state corresponding to the target object in the target frame.

[00209] In an embodiment, the state recognition module 1110 is further configured to:

[00210] input the target frame and each interaction information group into a trained interaction model when the region overlap ratio between the item region corresponding to the target item and the target region is less than or equal to the target overlap ratio threshold, to obtain the target object state corresponding to the target object in the target frame.

[00211] In an embodiment, the state recognition module 1110 is further configured to:

[00212] input the target frame and each interaction information group into the trained interaction model;

[00213] separately determine an overlapping region in the target frame based on each interaction information group;

[00214] extract an image feature of the overlapping region, to obtain a first interaction feature, obtain a second interaction feature based on an item and a target action that are in an interaction information group corresponding to the overlapping region, and extract an image feature of the target frame, to obtain a third interaction feature;

[00215] obtain a comprehensive interaction feature based on the first interaction feature, the second interaction feature, and the third interaction feature; and

[00216] output, based on the comprehensive interaction feature, the target object state corresponding to the target object in the target frame.

[00217] In an embodiment, the object state recognition apparatus is further configured to:

[00218] obtain a second training image corresponding to a training action of a training object, and a training object state and a training information group that correspond to the second training image, the training object state being an object state corresponding to the training object in the second training image, and the training information group being determined based on position information of a region in which the training object is located in the second training image, position information of a region in which an item is located in a background image corresponding to the second training image and item information, and the training action;

[00219] input the second training image and the training information group into a to-be-trained interaction model, to obtain a predicted object state corresponding to the second training image; and

[00220] adjust a model parameter of the to-be-trained interaction model based on a difference between the training object state and the predicted object state, until a second convergence condition is met, to obtain the trained interaction model.

[00221] In an embodiment, the target video is a video stream obtained by capturing an image of a living environment of the target object in real time. The object state recognition apparatus is further configured to:

[00222] when the target object state is an abnormal object state, transmit an alarm message for the abnormal object state to a management terminal corresponding to the living environment.

[00223] In an embodiment, the target video is the video stream obtained by capturing the image of the living environment of the target object in real time. The object state recognition apparatus is further configured to:

[00224] obtain an adjacent frame set of the target frame from the target video when the target object state is the abnormal object state;

[00225] perform target object recognition on an adjacent frame in the adjacent frame set, and use an adjacent frame in which the target object is recognized as an auxiliary frame;

[00226] determine an auxiliary object from the auxiliary frame, the auxiliary object and the target object in the target frame being different objects of the same type; and

[00227] transmit the alarm message for the abnormal object state to an object terminal corresponding to the auxiliary object.

[00228] In an embodiment, the target video is a video stream obtained by capturing, by a target device arranged in the living environment of the target object, the image of the living environment in real time. The object state recognition apparatus is further configured to:

[00229] when the target object state is the abnormal object state, transmit the alarm message for the abnormal object state to the target device, the target device being configured to: display an assistance requirement confirmation message for the abnormal object state, determine a target assistance manner for the target object in response to an assistance acceptance consent event for the assistance requirement confirmation message, and transmit an assistance message for the target object to an assistance device corresponding to the target assistance manner.

[00230] In an embodiment, the assistance requirement confirmation message includes a candidate assistance manner, and the target device is further configured to determine the target assistance manner in response to a selection operation for the candidate assistance manner.

[00231] In an embodiment, when the target assistance manner is a robot assistance manner, the assistance device is a robot whose distance from the living environment is within a preset threshold.

[00232] In the foregoing object state recognition apparatus, the target action of the target object in the target frame is analyzed based on the background frame and the target frame, to recognize the target object state corresponding to the target object in the target frame. The background frame can provide information about an item, and the target frame can provide information about the target object. An interaction information group is established based on the information about the item and the information about the target object, and an interaction relationship between the target object and the item is analyzed by using the interaction information group, so that accuracy of object state recognition can be effectively improved. When an object state is recognized, an interaction relationship between the target object and an environment is

considered, so that object states that are easily confused due to the same action can be effectively distinguished, thereby improving accuracy of object state recognition.

[00233] Each module in the foregoing object state recognition apparatus may be implemented entirely or partially through software, hardware, and a combination thereof. The foregoing modules may be embedded in or independent of a processor in the computer device in a form of hardware, or may be stored in a memory in the computer device in a form of software, so that the processor invokes and performs operations corresponding to the foregoing modules.

[00234] In an embodiment, a computer device is provided. The computer device may be a server, and an internal structural diagram thereof may be as shown in FIG. 12. The computer device includes a processor, a memory, an input/output (I/O) interface, and a communication interface. The processor, the memory, and the input/output interface are connected through a system bus, and the communication interface is connected to the system bus through the input/output interface. The processor of the computer device is configured to provide computing and control capabilities. The memory of the computer device includes a non-volatile storage medium and an internal memory. The non-volatile storage medium stores an operating system, computer-readable instructions and a database. The internal memory provides an environment for the operating system and the computer-readable instructions in the non-volatile storage medium. The database of the computer device is configured to store data such as various models and various reference information. The input/output interface of the computer device is configured to exchange information between the processor and an external device. The communication interface of the computer device is configured to connect to and communicate with an external terminal through a network. The computer-readable instructions, when executed by the processor, implement an object state recognition method.

[00235] In an embodiment, a computer device is provided. The computer device may be a terminal, and an internal structural diagram thereof may be as shown in FIG. 13. The computer device includes a processor, a memory, an input/output interface, a communication interface, a display unit, and an input apparatus. The processor, the memory, and the input/output interface are connected through a system bus, and the communication interface, the display unit, and the input apparatus are connected to the system bus through the input/output interface. The processor of the computer device is configured to provide computing and control capabilities. The memory of the computer device includes a non-volatile storage medium and an internal memory. The non-volatile storage medium stores an operating system and computer-readable instructions. The internal memory provides an environment for the operating system and the computer-readable instructions

in the non-volatile storage medium. The input/output interface of the computer device is configured to exchange information between the processor and an external device. The communication interface of the computer device is configured to perform wired or wireless communication with an external terminal. The wireless communication may be implemented through Wi-Fi, a mobile cellular network, near-field communication (NFC), or another technology. The computer-readable instructions, when executed by the processor, implement an object state recognition method. The display unit of the computer device is configured to form a visually visible image, and may be a display screen, a projection apparatus, or a virtual reality imaging apparatus. The display screen may be a liquid crystal display screen or an e-ink display screen. The input apparatus of the computer device may be a touch layer covering the display screen, or may be a button, a trackball, or a touchpad disposed on a housing of the computer device, or may be an external keyboard, touchpad, a mouse or the like.

[00236] A person skilled in the art may understand that the structures shown in FIG. 12 and FIG. 13 are only block diagrams of a partial structure related to a solution in this application, and FIG. 12 and FIG. 13 do not constitute a limitation to the computer device to which the solution in this application is applied. Specifically, the computer device may include more components or fewer components than those shown in FIG. 12 and FIG. 13, or some components may be combined, or a different component deployment may be used.

[00237] In an embodiment, a computer device is further provided. The computer device includes a memory and a processor. The memory stores computer-readable instructions, and the processor, when executing the computer-readable instructions, implements operations in the foregoing method embodiments.

[00238] In an embodiment, a computer-readable storage medium is provided. The computer-readable storage medium stores computer-readable instructions. The computer-readable instructions, when executed by a processor, implement operations in the foregoing method embodiments.

[00239] In an embodiment, a computer program product is provided. The computer program product includes computer-readable instructions. The computer-readable instructions, when executed by a processor, implement operations in the foregoing method embodiments.

[00240] User information (including, but not limited to, user equipment information, user personal information, and the like) and data (including, but not limited to, data for analysis, stored data, displayed data, and the like) involved in this application are information and data authorized

by a user or sufficiently authorized by all parties. Related data needs to be collected, used, and processed in accordance with related regulations.

[00241] A person of ordinary skill in the art may understand that some or all procedures in the method in the foregoing embodiments may be implemented by a computer-readable instruction instructing related hardware, the computer-readable instruction may be stored in a non-volatile computer readable storage medium, and when the computer-readable instruction is executed, the procedures in the foregoing method embodiments may be implemented. Any reference to a memory, a database or another medium used in the embodiments provided in this application may include at least one of a non-volatile or volatile memory. The non-volatile memory may include a read-only memory (ROM), a magnetic tape, a floppy disk, a flash memory, an optical memory, a high-density embedded non-volatile memory, a resistive random access memory (ReRAM), a magnetoresistive random access memory (MRAM), a ferroelectric random access memory (FRAM), a phase change memory (PCM), a graphene memory, and the like. The volatile memory may include a random access memory (RAM), an external high-speed cache, or the like. By way of illustration rather than limitation, the RAM may be in various forms, such as a static random access memory (SRAM) or a dynamic random access memory (DRAM). The database involved in the embodiments provided in this application may include at least one of a relational database or a non-relational database. The non-relational database may include a distributed database based on a block chain, and is not limited thereto. The processor involved in the embodiments provided in this application may be a general purpose processor, a central processing unit, a graphics processing unit, a digital signal processor, a programmable logic unit, a data processing logic unit based on quantum computing, and the like, and is not limited thereto.

[00242] The technical features of the foregoing embodiments may be randomly combined. For concise description, not all possible combinations of the technical features in the foregoing embodiments are described, but as long as combinations of the technical features do not conflict each other, the combinations of the technical features are to be considered as falling within the scope of this specification.

[00243] The foregoing embodiments show only several implementations of this application, and descriptions thereof are in detail, but shall not be understood as limiting the patent scope of this application. A person of ordinary skill in the art may further make variations and improvements without departing from the ideas of this application, which shall fall within the protection scope of this application. Therefore, the protection scope of this application shall be subject to the appended claims.

CLAIMS

1. An object state recognition method, performed by a computer device, and the method comprising:

obtaining a target video;

obtaining a background frame corresponding to the target video, and determining, from the target video, a target frame corresponding to a target action of a target object;

performing image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame;

generating an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item; and

determining, based on each interaction information group, a target object state corresponding to the target object in the target frame.

2. The method according to claim 1, wherein the obtaining a background frame corresponding to the target video comprises:

performing target object recognition on a video frame in the target video, and each time it is found that no target object is recognized in a target quantity of consecutive video frames, obtaining a last video frame from the target quantity of consecutive video frames as the background frame.

3. The method according to claim 2, wherein the performing target object recognition on a video frame in the target video comprises:

performing object recognition on the video frame in the target video based on a general object recognition algorithm, and using a first recognized living object as the target object;

determining a target object recognition algorithm based on a protection priority of the target object; and

performing target object recognition on the video frame in the target video based on the target object recognition algorithm.

4. The method according to claim 2 or 3, wherein the target quantity is negatively correlated

with recognition accuracy corresponding to a used target object recognition algorithm.

5. The method according to any one of claims 1 to 3, wherein the determining, from the target video, a target frame corresponding to a target action of a target object comprises:

performing target object recognition on the video frame in the target video, and using a video frame in which the target object is recognized as a candidate frame;

obtaining candidate frames within a target time period to form a candidate frame sequence;

performing target object skeleton point recognition on candidate frames in the candidate frame sequence, to obtain skeleton point data corresponding to the candidate frames, and forming a skeleton point data sequence corresponding to the candidate frame sequence;

inputting the skeleton point data sequence into a multi-frame action recognition model, to obtain a predicted action corresponding to the candidate frame sequence; and

when the predicted action is used as the target action, determining, from the candidate frame sequence, the target frame corresponding to the target action of the target object.

6. The method according to claim 5, wherein the determining, from the candidate frame sequence, the target frame corresponding to the target action of the target object comprises:

separately inputting each candidate frame in the candidate frame sequence into a trained single-frame action recognition model, to obtain an action recognition result corresponding to each candidate frame, the action recognition result comprising a quantity and actions of target objects in each candidate frame; and

obtaining a candidate frame whose action recognition result comprises a single target object and the target action, and using the candidate frame as the target frame corresponding to the target action of the target object.

7. The method according to claim 6, wherein the method further comprises:

obtaining a first training image, and a target object position training label and a target object action training label that correspond to the first training image;

inputting the first training image into a to-be-trained single-frame action recognition model, to obtain a target object position prediction label and a target object action prediction label that correspond to the first training image; and

adjusting a model parameter of the to-be-trained single-frame action recognition model based on a difference between the target object position training label and the target object position

prediction label and a difference between the target object action training label and the target object action prediction label, until a first convergence condition is met, to obtain the trained single-frame action recognition model.

8. The method according to any one of claims 1 to 7, wherein the determining, based on each interaction information group, a target object state corresponding to the target object in the target frame comprises:

obtaining reference object states corresponding to the target object under the target action, and obtaining reference items respectively corresponding to the reference object states;

using a reference item in the background frame as a target item, and obtaining a reference overlap ratio threshold corresponding to a reference object state corresponding to the target item as a target overlap ratio threshold; and

for an interaction information group corresponding to the target item, when a region overlap ratio between an item region corresponding to the target item and the target region is greater than the target overlap ratio threshold, using the reference object state corresponding to the target item as the target object state corresponding to the target object in the target frame.

9. The method according to claim 8, wherein the method further comprises:

inputting the target frame and each interaction information group into a trained interaction model when the region overlap ratio between the item region corresponding to the target item and the target region is less than or equal to the target overlap ratio threshold, to obtain the target object state corresponding to the target object in the target frame.

10. The method according to any one of claims 1 to 7, wherein the determining, based on each interaction information group, a target object state corresponding to the target object in the target frame comprises:

inputting the target frame and each interaction information group into a trained interaction model;

separately determining an overlapping region in the target frame based on each interaction information group;

extracting an image feature of the overlapping region, to obtain a first interaction feature, obtaining a second interaction feature based on an item and a target action that are in an interaction information group corresponding to the overlapping region, and extracting an image feature of the target frame, to obtain a third interaction feature;

obtaining a comprehensive interaction feature based on the first interaction feature, the second interaction feature, and the third interaction feature; and

outputting, based on the comprehensive interaction feature, the target object state corresponding to the target object in the target frame.

11. The method according to claim 10, wherein the method further comprises:

obtaining a second training image corresponding to a training action of a training object, and a training object state and a training information group that correspond to the second training image, the training object state being an object state corresponding to the training object in the second training image, and the training information group being determined based on position information of a region in which the training object is located in the second training image, position information of a region in which an item is located in a background image corresponding to the second training image and item information, and the training action;

inputting the second training image and the training information group into a to-be-trained interaction model, to obtain a predicted object state corresponding to the second training image; and

adjusting a model parameter of the to-be-trained interaction model based on a difference between the training object state and the predicted object state, until a second convergence condition is met, to obtain the trained interaction model.

12. The method according to any one of claims 1 to 11, wherein the target video is a video stream obtained by capturing an image of a living environment of the target object in real time, and the method further comprises:

transmitting, when the target object state is an abnormal object state, an alarm message for the abnormal object state to a management terminal corresponding to the living environment.

13. The method according to any one of claims 1 to 11, wherein the target video is a video stream obtained by capturing an image of a living environment of the target object in real time, and the method further comprises:

obtaining an adjacent frame set of the target frame from the target video when the target object state is an abnormal object state;

performing target object recognition on an adjacent frame in the adjacent frame set, and using an adjacent frame in which the target object is recognized as an auxiliary frame;

determining an auxiliary object from the auxiliary frame, the auxiliary object and the target object in the target frame being different objects of the same type; and

transmitting an alarm message for the abnormal object state to an object terminal corresponding to the auxiliary object.

14. The method according to any one of claims 1 to 11, wherein the target video is a video stream obtained by capturing, by a target device arranged in a living environment of the target object, an image of the living environment in real time, and the method further comprises:

transmitting, when the target object state is an abnormal object state, an alarm message for the abnormal object state to the target device, the target device being configured to: display an assistance requirement confirmation message for the abnormal object state, determine a target assistance manner for the target object in response to an assistance acceptance consent event for the assistance requirement confirmation message, and transmit an assistance message for the target object to an assistance device corresponding to the target assistance manner.

15. The method according to claim 14, wherein the assistance requirement confirmation message comprises a candidate assistance manner, and the target device is further configured to determine the target assistance manner in response to a selection operation for the candidate assistance manner.

16. The method according to claim 14, wherein when the target assistance manner is a robot assistance manner, the assistance device is a robot whose distance from the living environment is within a preset threshold.

17. An object state recognition apparatus, the apparatus comprising:

a video obtaining module, configured to obtain a target video;

an image determining module, configured to: obtain a background frame corresponding to the target video, and determine, from the target video, a target frame corresponding to a target action of a target object;

an image segmentation module, configured to perform image segmentation on the background frame and the target frame, to obtain an item region corresponding to each item in the background frame and a target region corresponding to the target object in the target frame;

an information group generation module, configured to generate an interaction information group based on position information of the target region and position information of the item region, the item corresponding to the item region, and the target action, to obtain an interaction information group between the target object and each item; and

a state recognition module, configured to determine, based on each interaction information

group, a target object state corresponding to the target object in the target frame.

18. A computer device, comprising a memory and a processor, the memory storing computer-readable instructions, and the processor, when executing the computer-readable instructions, implementing operations of the method according to any one of claims 1 to 16.

19. A computer-readable storage medium, having computer-readable instructions stored therein, the computer-readable instructions, when executed by a processor, implementing operations of the method according to any one of claims 1 to 16.

20. A computer program product, comprising computer-readable instructions, the computer-readable instructions, when executed by a processor, implementing operations of the method according to any one of claims 1 to 16.