

OBJECT SEQUENCE RECOGNITION METHOD, NETWORK TRAINING METHOD, APPARATUSES, DEVICE, AND MEDIUM

TECHNICAL FIELD

[0001] Embodiments of the application relate to the technical field of image processing, and relate, but not limited, to an object sequence recognition method, a network training method, apparatuses, a device, and a medium.

BACKGROUND

[0002] Sequence recognition on an image is an important research subject in computer vision. A sequence recognition algorithm is widely applied to scene text recognition, license plate recognition and other scenes. In the related art, a neural network is used to recognize an image including sequential objects. The neural network may be obtained by training taking classes of objects in sequential objects as supervision information.

[0003] In the related art, an effect of performing sequence recognition on an object sequence in an image by a common sequence recognition method is not so good.

SUMMARY

[0004] The embodiments of the application provide technical solutions to the recognition of an object sequence.

[0005] The technical solutions of the embodiments of the application are implemented as follows.

[0006] An embodiment of the application provides an object sequence recognition method, which may include the following operations.

- [0007] An image including an object sequence is acquired.
- [0008] Feature extraction is performed on the image including the object sequence using an object sequence recognition network to obtain a feature sequence. Herein, supervision information in a training process of the object sequence recognition network may at least include first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group may include at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group may satisfy a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group may satisfies a preset consistency condition.
- [0009] A class of each object in the object sequence is determined based on the feature sequence.
- [0010] In some embodiments, the operation that feature extraction is performed on the image including the object sequence using an object sequence recognition network to obtain a feature sequence may include the following operations. Feature extraction is performed on the image including the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map. The feature map is split to obtain the feature sequence. As such, it is easy to subsequently recognize the classes of the objects in the feature sequence more accurately.
- [0011] In some embodiments, the operation that feature extraction is performed on the image including the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map may include the following operations. The image including the object sequence is down-sampled using the convolutional subnetwork in a length dimension of the image including the object sequence in a first direction to obtain a first-dimensional feature, the first direction being different from an arrangement direction of the objects in the object sequence. A feature in a length dimension of the image including the object sequence in a second direction is extracted based on a length of the image including the object sequence in the second direction to obtain a second-dimensional feature. The feature map is obtained based on

the first-dimensional feature and the second-dimensional feature. As such, feature information of the image including the object sequence in the dimension in the second direction may be maximally retained.

[0012] In some embodiments, the operation that the feature map is split to obtain the feature sequence may include the following operations. The feature map is pooled in the first direction to obtain a pooled feature map. The pooled feature map is split in the second direction to obtain the feature sequence. Accordingly, the feature map is split in the second direction after being pooled in the first direction, so that the feature sequence may include more detail information of the image including the object sequence in the second direction.

[0013] An embodiment of the application provides a method for training an object sequence recognition network, which may include the following operations. A sample image group is acquired. Herein, the sample image group may include at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group may satisfy a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group may satisfies a preset consistency condition, and each frame of sample image may include class labeling information of a sample object sequence.

[0014] The sample image group is input to an object sequence recognition network to be trained, and feature extraction is performed to obtain sample feature sequences.

[0015] Class prediction is performed on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group.

[0016] A first loss and a second loss set are determined based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group. Herein, the first loss may be negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images may be determined based on sample feature sequences of the multiple frames of different sample images and/or predicted

classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set may be configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence.

[0017] A network parameter of the object sequence recognition network to be trained is adjusted according to the first loss and the second loss set such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition. Accordingly, the second loss set for supervising object sequences and the first loss for supervising similarities between images in a group of sample images are introduced to a training process, so that the accuracy of recognizing a class of each object in an image may be improved.

[0018] In some embodiments, the operation that a sample image group is acquired may include the following operations. A sample video stream including the sample object sequence is acquired. Sample object sequence detection is performed on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images. At least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images are determined to form the sample image group. As such, the richness of sample image group data may be improved.

[0019] In some embodiments, the operation that the sample image group is input to an object sequence recognition network to be trained and feature extraction is performed to obtain sample feature sequences may include the following operations. Feature extraction is performed on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group. The sample feature map of each sample image in the sample image group is split to obtain the sample feature sequence of each sample image in the sample image group. As such, the obtained sample feature sequence may retain more features in a second direction to facilitate the improvement of the accuracy of subsequently recognizing classes of sample objects in

the sample feature sequence.

[0020] In some embodiments, the operation that feature extraction is performed on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group may include the following operations. Each sample image in the sample image group is down-sampled using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature, the first direction being different from an arrangement direction of sample objects in the sample object sequence. A feature in a length dimension of each sample image in the sample image group in a second direction is extracted based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature. The sample feature map of each sample image in the sample image group is obtained based on the first-dimensional sample feature and the second-dimensional sample feature. As such, feature information in the dimension of each sample image in the sample image group in the second direction may be maximally retained.

[0021] In some embodiments, the operation that the sample feature map of each sample image in the sample image group is split to obtain the sample feature sequence of each sample image in the sample image group may include the following operations. The sample feature map of each sample image in the sample image group is pooled in the first direction to obtain a pooled sample feature map of each sample image in the sample image group. The pooled sample feature map of each sample image in the sample image group is split in the second direction to obtain the sample feature sequence of each sample image in the sample image group. As such, the obtained sample feature sequence may retain more features in the second direction to make it easy to subsequently recognize classes of the sample objects in the sample feature sequence more accurately.

[0022] In some embodiments, the operation that a network parameter of the object sequence recognition network to be trained is adjusted according to the first loss and the second loss set such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition may include the

following operations. Weighted fusion is performed on the first loss and the second loss set to obtain a total loss. The network parameter of the object sequence recognition network to be trained is adjusted according to the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition. Accordingly, two loss functions are fused as the total loss, and the network is trained using the total loss, so that the object recognition performance of the network may be improved.

[0023] In some embodiments, the operation that weighted fusion is performed on the first loss and the second loss set to obtain a total loss may include the following operations. The first loss is adjusted using a first preset weight to obtain a third loss. A class supervision weight is determined based on the number of the sample images in the sample image group, multiple different sample images in the same sample image group corresponding to the same class supervision weight. The second losses in the second loss set are fused based on the class supervision weight and a second preset weight to obtain a fourth loss. The total loss is determined based on the third loss and the fourth loss. Accordingly, the object sequence recognition network to be trained is trained using the total loss obtained by fusing the third loss and the fourth loss, so that a prediction effect of the whole network may be improved, and an object recognition network with relatively high performance may be obtained.

[0024] In some embodiments, the operation that the second losses in the second loss set are fused based on the class supervision weight and a second preset weight to obtain a fourth loss may include the following operations. The class supervision weight is assigned to each second loss in the second loss set to obtain an updated loss set including at least two updated losses. The updated losses in the updated loss set are fused to obtain a fused loss. The fused loss is adjusted using the second preset weight to obtain the fourth loss. Accordingly, Connectionist Temporal Classification (CTC) losses of prediction results of each sample image in a group of sample images are fused in the training process, so that the performance of the trained recognition network may be improved.

[0025] An embodiment of the application provides an object sequence recognition apparatus, which may include a first acquisition module, a first extraction module, and a

first determination module.

[0026] The first acquisition module may be configured to acquire an image including an object sequence.

[0027] The first extraction module may be configured to perform feature extraction on the image including the object sequence using an object sequence recognition network to obtain a feature sequence. Supervision information in a training process of the object sequence recognition network may at least include first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image. Each sample image group may include at least two frames of sample images extracted from the same video stream. Timing of each frame of sample image in each sample image group may satisfy a preset timing condition. Positions of a sample object sequence in the frames of sample image in a sample image group may satisfy a preset consistency condition.

[0028] The first determination module may be configured to determine a class of each object in the object sequence based on the feature sequence.

[0029] In some embodiments, the first extraction module may include a first feature extraction submodule and a first splitting submodule. The first feature extraction submodule may be configured to perform feature extraction on the image including the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map. The first splitting submodule may be configured to split the feature map to obtain the feature sequence.

[0030] In some embodiments, the first feature extraction submodule may include a first down-sampling subunit, a first feature extraction subunit, and a first feature map determination subunit. The first down-sampling subunit may be configured to down-sample the image including the object sequence using the convolutional subnetwork in a length dimension of the image including the object sequence in a first direction to obtain a first-dimensional feature, the first direction being different from an arrangement direction of the objects in the object sequence. The first feature extraction subunit may be

configured to extract a feature in a length dimension of the image including the object sequence in a second direction based on a length of the image including the object sequence in the second direction to obtain a second-dimensional feature. The first feature map determination subunit may be configured to obtain the feature map based on the first-dimensional feature and the second-dimensional feature.

[0031] In some embodiments, the first splitting submodule may include a first pooling subunit and a first splitting subunit. The first pooling subunit may be configured to pool the feature map in the first direction to obtain a pooled feature map. The first splitting subunit may be configured to split the pooled feature map in the second direction to obtain the feature sequence.

[0032] An embodiment of the application provides an apparatus for training an object sequence recognition network, which may include a second acquisition module, a second extraction module, a second prediction module, a second determination module, and a first adjustment module.

[0033] The second acquisition module may be configured to acquire a sample image group. The sample image group may include at least two frames of sample images extracted from the same video stream. Timing of each frame of sample image in each sample image group may satisfy a preset timing condition. Positions of a sample object sequence in the frames of sample image in a sample image group may satisfy a preset consistency condition. Each frame of sample image may include class labeling information of a sample object sequence.

[0034] The second extraction module may be configured to input the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain sample feature sequences.

[0035] The second prediction module may be configured to perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group.

[0036] The second determination module may be configured to determine a first loss

and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group. The first loss may be negatively correlated with similarities between multiple frames of different sample images in the sample images. The similarities between the multiple frames of different sample images may be determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images. A second loss in the second loss set may be configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence.

[0037] The first adjustment module may be configured to adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.

[0038] In some embodiments, the second acquisition module may include a second acquisition submodule, a second detection submodule, and a second forming submodule. The second acquisition submodule may be configured to acquire a sample video stream including the sample object sequence. The second detection submodule may be configured to perform sample object sequence detection on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images. The second forming submodule may be configured to determine at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images to form the sample image group.

[0039] In some embodiments, the second extraction module may include a second feature extraction submodule and a second splitting submodule. The second feature extraction submodule may be configured to perform feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each

sample image in the sample image group. The second splitting submodule may be configured to split the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the sample image group.

[0040] In some embodiments, the second feature extraction submodule may include a second down-sampling subunit, a second feature extraction subunit, and a second feature map determination subunit. The second down-sampling subunit may be configured to down-sample each sample image in the sample image group using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature, the first direction being different from an arrangement direction of sample objects in the sample object sequence. The second feature extraction subunit may be configured to extract a feature in a length dimension of each sample image in the sample image group in a second direction based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature. The second feature map determination subunit may be configured to obtain the sample feature map of each sample image in the sample image group based on the first-dimensional sample feature and the second-dimensional sample feature.

[0041] In some embodiments, the second splitting submodule may include a second pooling subunit and a second splitting subunit. The second pooling subunit may be configured to pool the sample feature map of each sample image in the sample image group in the first direction to obtain a pooled sample feature map of each sample image in the sample image group. The second splitting subunit may be configured to split the pooled sample feature map of each sample image in the sample image group in the second direction to obtain the sample feature sequence of each sample image in the sample image group.

[0042] In some embodiments, the first adjustment module may include a fusion submodule and an adjustment submodule. The fusion submodule may be configured to perform weighted fusion on the first loss and the second loss set to obtain a total loss. The adjustment submodule may be configured to adjust the network parameter of the object sequence recognition network to be trained according to the total loss such that the

loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.

[0043] In some embodiments, the fusion submodule may include a first adjustment unit, a weight determination unit, a fusion unit, and a determination unit. The first adjustment unit may be configured to adjust the first loss using a first preset weight to obtain a third loss. The weight determination unit may be configured to determine a class supervision weight based on the number of the sample images in the sample image group, multiple different sample images in the same sample image group corresponding to the same class supervision weight. The fusion unit may be configured to fuse the second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss. The determination unit may be configured to determine the total loss based on the third loss and the fourth loss.

[0044] In some embodiments, the fusion unit may include an assignment subunit, a fusion subunit, and an adjustment subunit. The assignment subunit may be configured to assign the class supervision weight to each second loss in the second loss set to obtain an updated loss set including at least two updated losses. The fusion subunit may be configured to fuse the updated losses in the updated loss set to obtain a fused loss. The adjustment subunit may be configured to adjust the fused loss using the second preset weight to obtain the fourth loss.

[0045] An embodiment of the application provides a computer device, which may include a memory and a processor. A computer-executable instruction may be stored in the memory. The processor may run the computer-executable instruction in the memory to implement the abovementioned object sequence recognition method. Alternatively, the processor may run the computer-executable instruction in the memory to implement the abovementioned method for training an object sequence recognition network.

[0046] An embodiment of the application provides a computer storage medium, in which a computer-executable instruction may be stored. The computer-executable instruction may be executed to implement the abovementioned object sequence recognition method. Alternatively, the computer-executable instruction may be executed to implement the abovementioned method for training an object sequence recognition

network.

[0047] According to the object sequence recognition method, network training method, apparatuses, device and medium provided in the embodiments of the application, feature extraction is performed on the image including the object sequence at first using the object sequence recognition network including the first supervision information for supervising the similarity between the at least two frames of different sample images extracted from the same video stream in the sample image group and the second supervision information for supervising the class of the sample object sequence in each sample image group to obtain the feature sequence. Then, the class of each object in the object sequence is determined based on the feature sequence. As such, the consistency of feature extraction and recognition results of similar images obtained by the object sequence recognition network is improved, relatively high robustness is achieved, and the object sequence recognition accuracy is improved.

BRIEF DESCRIPTION OF THE DRAWINGS

[0048] In order to describe the technical solutions of the embodiments of the application more clearly, the drawings required to be used in the descriptions about the embodiments will be simply introduced below. It is apparent that the drawings described below are merely some embodiments of the application. Other drawings may further be obtained by those of ordinary skill in the art according to these drawings without creative work.

[0049] FIG. 1 is an implementation flowchart of a first object sequence recognition method according to an embodiment of the application.

[0050] FIG. 2 is an implementation flowchart of a second object sequence recognition method according to an embodiment of the application.

[0051] FIG. 3 is an implementation flowchart of a method for training an object sequence recognition network according to an embodiment of the application.

[0052] FIG. 4 is a structure diagram of an object sequence recognition network according to an embodiment of the application.

[0053] FIG. 5 is a schematic diagram of an application scene of an object sequence recognition network according to an embodiment of the application.

[0054] FIG. 6A is a structure composition diagram of an object sequence recognition apparatus according to an embodiment of the application.

[0055] FIG. 6B is a structure composition diagram of an apparatus for training an object sequence recognition network according to an embodiment of the application.

[0056] FIG. 7 is a composition structure diagram of a computer device according to an embodiment of the application.

DETAILED DESCRIPTION

[0057] In order to make the purposes, technical solutions, and advantages of the embodiments of the application clearer, specific technical solutions of the disclosure will further be described below in combination with the drawings in the embodiments of the application in detail. The following embodiments are adopted to describe the application rather than limit the scope of the application.

[0058] "Some embodiments" involved in the following descriptions describes a subset of all possible embodiments. However, it can be understood that "some embodiments" may be the same subset or different subsets of all the possible embodiments, and may be combined without conflicts.

[0059] Term "first/second/third" involved in the following descriptions is only for distinguishing similar objects, and does not represent a specific sequence of the objects. It can be understood that "first/second/third" may be interchanged to specific sequences or orders if allowed to implement the embodiments of the application described herein in sequences except the illustrated or described ones.

[0060] Unless otherwise defined, all technical and scientific terms used herein have the same meanings as commonly understood by those skilled in the art of the application. Terms used in the application are only adopted to describe the embodiments of the application and not intended to limit the application.

[0061] Nouns and terms involved in the embodiments of the application will be described before the embodiments of the application are further described in detail. The nouns and terms involved in the embodiments of the application are suitable to be explained as follows.

[0062] 1) Deep Learning (DL) is introduced to Machine Learning (ML) as a new research direction in the field of ML to make ML closer to the initial goal: Artificial Intelligence (AI). DL refers to learning inherent laws and representation layers of sample data. Information obtained in these learning processes helps greatly the interpretation of data such as texts, images and sounds. The final goal of DL is to make machines able to analyze and learn like humans and able to recognize data such as texts, images and sounds.

[0063] 2) Pair loss: Paired samples are used for loss calculation in many metric learning methods in DL. For example, in a model training process, two samples are randomly selected, and a model is used to extract features and calculate a distance between the features of the two samples. If the two samples belong to the same class, the distance between the two samples is expected to be as short as possible, even 0. If the two samples belong to different classes, the distance between the two samples is expected to be as long as possible, even infinitely long. Various types of feature pair losses are derived based on this principle. These losses are used to calculate distances of sample pairs, and the model is updated by various optimization methods according to generated losses.

[0064] 3) CTC calculates a loss value, has the main advantage that unaligned data may be aligned automatically, and is mainly used for the training of sequential data that is not aligned in advance, e.g., voice recognition and Optical Character Recognition (OCR). In the embodiments of the application, a CTC loss may be used to supervise an overall prediction condition of a sequence during the early training of a network.

[0065] An exemplary application of an object sequence recognition device provided in the embodiments of the application will be described below. The device provided in the embodiments of the application may be implemented as various types of user terminals with an image collection function, such as a notebook computer, a tablet computer, a desktop computer, a camera, and a mobile device (e.g., a personal digital assistant, a dedicated messaging device, and a portable game device), or may be implemented as a server. The exemplary application of the device implemented as the terminal or the server will be described below.

[0066] A method may be applied to a computer device. A function realized by the method may be realized by a processor in the computer device by calling a program code. Of course, the program code may be stored in a computer storage medium. It can be seen that the computer device at least includes the processor and the storage medium.

[0067] An embodiment of the application provides an object sequence recognition method. As shown in FIG. 1, descriptions will be made in combination with the operations shown in FIG. 1.

[0068] In S101, an image including an object sequence is acquired.

[0069] In some embodiments, the object sequence may be a sequence formed by sequentially arranging any objects. A specific object type is not specially limited. The image including the object sequence may be an image including appearance information of the object sequence. The image including the object sequence may be an image collected by any electronic device with an image collection function, or may be an image acquired from another electronic device or a server.

[0070] In some embodiments, the image including the object sequence is at least one frame of image. The at least one frame of image may be an image of which timing satisfies a preset timing condition and in which a position of the same object sequence satisfies a preset consistency condition. In addition, the at least one frame of image may be a preprocessed image, e.g., an image obtained by image size unification and/or image pixel value unification.

[0071] In some embodiments, the image including the object sequence may be an

image collected in a game scene, and the object sequence may be tokens in a game in a game place, etc. Alternatively, the image including the object sequence is an image collected in a scene that planks of various materials or colors are stacked, and the object sequence may be a pile of stacked planks. Alternatively, the image including the object sequence is an image collected in a book stacking scene, and the object sequence may be a pile of stacked books.

[0072] In some possible implementation modes, an acquired video stream is preprocessed to obtain the image including the object sequence. That is, S101 may be implemented through the following process.

[0073] In a first step, a video stream including at least one object sequence is acquired.

[0074] In some embodiments, the video stream including the at least one object sequence may be collected by any electronic device with a video collection function. The video stream may include two or more image frames. Position information of the object sequence in each frame of image in the video stream may be the same or different. In addition, each frame of image in the video stream may be continuous or discontinuous in timing.

[0075] In a second step, an image parameter of a video frame is preprocessed according to a preset image parameter to obtain the image including the object sequence.

[0076] In some embodiments, the preset image parameter may be a preset image size parameter and/or a preset image pixel parameter. For example, the preset image parameter is a preset image width and a preset aspect ratio. In such case, a width of each frame of image in the video stream may be adjusted to the preset image width in a unified manner, and a height of each frame of image in the video stream may be adjusted according to the ratio. Meanwhile, for an image with an insufficient height, an image region that does not reach a preset height is filled with pixels, thereby obtaining an image including the object sequence. A pixel value for pixel filling may be determined as practically required. Alternatively, the preset image parameter is a preset image pixel parameter. In such case, a normalization operation is performed on image pixels of each frame of image in the video stream, for example, each pixel value of each frame of image

is scaled to interval (0, 1), to obtain an image including the object sequence.

[0077] In some embodiments, the image parameter of each frame of image in the video stream may be adjusted to obtain images of which image parameters are the same and which include the object sequence. Therefore, the probability that the image including the object sequence is deformed in a post-processing process may be reduced, and furthermore, the accuracy of recognizing the object sequence in a picture of the image including the object sequence may be improved.

[0078] In S102, feature extraction is performed on the image including the object sequence using an object sequence recognition network to obtain a feature sequence.

[0079] In some embodiments, supervision information in a training process of the object sequence recognition network at least includes first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image. Each sample image group includes at least two frames of sample images extracted from the same video stream. Timing of each frame of sample image in each sample image group satisfies a preset timing condition. A position of the same sample object sequence in each frame of sample image in a sample image group satisfies a preset consistency condition. The timing of the sample image may be a timing position thereof in the video stream or collection time of the sample image. That the timing of the sample images satisfies the preset timing condition may refer to that a distance between timing positions of the sample images in the video stream is less than a preset threshold, or may refer to that a distance between collection time of the sample images is less than a preset threshold.

[0080] In some embodiments, that the position of the same sample object sequence in each frame of sample image in a sample image group satisfies the preset consistency condition may refer to that the positions of the sample object sequence in a picture of each sample image in the sample image group are the same or similarities satisfy a preset threshold, or may refer to that regions of a detection box corresponding to the sample object sequence in each sample image in the sample image group are the same or similarities satisfy a preset threshold.

- [0081] In some embodiments, feature extraction is performed on the image including the object sequence using the object sequence recognition network to obtain the feature sequence. Each feature in the feature sequence may correspond to an object in the object sequence. Alternatively, multiple features in the feature sequence correspond to an object in the object sequence.
- [0082] In some embodiments, the image including the object sequence is input to the object sequence recognition network. Feature extraction may be performed on the image including the object sequence at first using a convolutional neural network part in the object sequence recognition network to obtain a feature map. Then, the feature map is split according to a certain manner, thereby splitting the feature map extracted by the convolutional neural network into a plurality of feature sequences. As such, subsequent classification of each object in the object sequence in the image including the object sequence is facilitated.
- [0083] In S103, a class of each object in the object sequence is determined based on the feature sequence.
- [0084] In some embodiments, class prediction is performed on each feature in the feature sequence to obtain a classification result of each feature in the feature sequence. Then, class information of each object in the at least one object sequence is determined based on the classification result of the feature sequence. The feature sequence includes multiple features. The classification result of each feature may be an object class corresponding to each feature.
- [0085] In some embodiments, the class of each object in the object sequence includes the class of each object and a sequence length of objects of the same class in the object sequence.
- [0086] In some embodiments, a class of the feature in the feature sequence is predicted using a classifier in the object sequence recognition network, thereby obtaining a predicted probability of the class of each object in the object sequence. The classification result of the feature sequence may represent a probability that the object sequence in the feature sequence belongs to a class corresponding to each classification

label. A class corresponding to a classification label, of which a probability value is greater than a certain threshold, in a group of probabilities corresponding to a feature sequence is determined as a class of an object corresponding to the feature in the feature sequence.

[0087] According to the object sequence recognition method provided in the embodiment of the application, feature extraction is performed on the image including the object sequence at first using the object sequence recognition network including the first supervision information for supervising the similarity between the at least two frames of sample images extracted from the same video stream in the sample image group and the second supervision information for supervising the class of the sample object sequence in each sample image group to obtain the feature sequence. Then, the class of each object in the object sequence is determined based on the feature sequence. As such, the consistency of feature extraction and recognition results of similar images obtained by the object sequence recognition network is improved, relatively high robustness is achieved, and the object sequence recognition accuracy is improved.

[0088] In some embodiments, the feature extraction of the image including the object sequence is implemented by a convolutional network obtained by finely adjusting the structure of a Residual Network (ResNet), thereby obtaining the feature sequence. That is, S102 may be implemented through the operations shown in FIG. 2. FIG. 2 is another implementation flowchart of an object sequence recognition method according to an embodiment of the application. The following descriptions will be made in combination with the operations shown in FIGS. 1 and 2.

[0089] In S201, feature extraction is performed on the image including the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map.

[0090] In some embodiments, the convolutional subnetwork in the object sequence recognition network may be a convolutional network obtained by fine adjustment based on a network structure of a ResNet. The convolutional subnetwork in the object sequence recognition network may be obtained by adjusting three layers of convolutional blocks in the ResNet into multiple blocks which are stacked into the same topology structure in

parallel, or may be obtained by changing convolutional layers 3 and 4 of which last strides are (2, 2) respectively in the ResNet into convolutional layers of which strides are (1, 2).

[0091] In some embodiments, a high-layer feature of the image including the object sequence may be extracted using the convolutional subnetwork in the object sequence recognition network, thereby obtaining the feature map. The high-layer feature may be relatively complex in the image including the object sequence, and is not a texture, color, edge, corner angle and other feature information in the image. For example, the high-layer feature may be golden hair and colorful flowers.

[0092] In some possible implementation modes, feature extraction is performed on the image including the object sequence in the object sequence recognition network, thereby obtaining a feature map of which a width is changed and a height is kept unchanged. That is, S201 may be implemented through the following S211 to S213 (not shown in the figure).

[0093] In S211, the image including the object sequence is down-sampled using the convolutional subnetwork in a length dimension of the image including the object sequence in a first direction to obtain a first-dimensional feature.

[0094] In some embodiments, the first direction is different from an arrangement direction of the objects in the object sequence. For example, if the object sequence is multiple objects arranged or stacked in a height direction, namely the arrangement direction of the objects in the object sequence is the height direction, the first direction may be a width direction of the object sequence. If the object sequence is multiple objects arranged in a horizontal direction, namely the arrangement direction of the objects in the object sequence is the horizontal direction, the first direction may be the height direction of the object sequence.

[0095] In some embodiments, strides in the first direction in last strides of convolutional layers 3 and 4 in the network structure of the ResNet are kept at 2 and unchanged, and a convolutional network obtained by adjusting the network structure of the ResNet is taken as the convolutional subnetwork in the object sequence recognition

network. In this manner, the image including the object sequence may be down-sampled in the length dimension of the first image in the first direction. That is, a length of the obtained feature map in the first direction is a half of a length of the image including the object sequence in the first direction. For example, the object sequence is multiple objects stacked in a height direction. In such case, width strides in the last strides of convolutional layers 3 and 4 in the network structure of the ResNet are kept at 2 and unchanged. In this manner, down-sampling in a width dimension of the image including the object sequence is implemented, and a width of the obtained feature map is changed to a half of a width of the first image.

[0096] In S212, a feature in a length dimension of the image including the object sequence in a second direction is extracted based on a length of the image including the object sequence in the second direction to obtain a second-dimensional feature.

[0097] In some embodiments, the second direction is the same as the arrangement direction of the objects in the object sequence. Strides in the second direction in the last strides of convolutional layers 3 and 4 in the network structure of the ResNet are changed from 2 to 1. In this manner, down-sampling is not performed in the length dimension of the image including the object sequence in the second direction, namely the length of the image including the object sequence in the second direction is kept. Meanwhile, feature extraction is performed in the length direction of the image including the object sequence in the second direction to obtain a second-dimensional feature the same as the length of the image including the object sequence in the second direction.

[0098] In S213, the feature map is obtained based on the first-dimensional feature and the second-dimensional feature.

[0099] In some embodiments, the first-dimensional feature of the image including the object sequence may be combined with the second-dimensional feature of the image including the object sequence to obtain the feature map of the image including the object sequence.

[00100] In some embodiments, the last strides of convolutional layers 3 and 4 in the ResNet are changed from (2, 1) to (1, 2), so that the image including the object sequence

is not down-sampled in the height dimension, and meanwhile, is down-sampled in the width dimension. As such, feature information of the image including the object sequence in the height dimension may be maximally retained.

[00101] In S202, the feature map is split to obtain the feature sequence.

[00102] In some embodiments, the feature map may be split based on dimension information of the feature map, thereby obtaining the feature sequence. The dimension information of the feature map includes a dimension in the first direction and a dimension in the second direction. For example, the dimension information is a height dimension and a width dimension. In such case, the feature map is split based on the height dimension and the width dimension, thereby obtaining the feature sequence of the image including the object sequence. The feature map may be split according to equal size information when being split based on the height dimension and the width dimension.

[00103] In some embodiments, the feature map is pooled at first in the dimension of the feature map in the first direction, and then a splitting operation is performed on the feature map in the dimension of the feature map in the second direction, thereby splitting the feature map into the feature sequence. In this manner, feature extraction is performed on the image including the object sequence using the object sequence recognition network obtained by training based on two loss functions to obtain the feature map, and the feature map is split according to the dimension information, so that the obtained feature sequence may retain more features in the second direction to make it easy to subsequently recognize the class of the object sequence in the feature sequence more accurately.

[00104] In some possible implementation modes, the feature map is pooled in the dimension in the first direction to obtain a pooled map, and the obtained pooled map is split in the dimension in the second direction to obtain the feature sequence. That is, S202 may be implemented through S221 and S222 (not shown in the figure).

[00105] In S221, the feature map is pooled in the first direction to obtain a pooled feature map.

[00106] In some embodiments, average pooling is performed on the feature map in the

dimension of the feature map in the first direction, and meanwhile, the dimension of the feature map in the second direction and a channel dimension are kept unchanged, to obtain the pooled feature map. For example, a dimension of the feature map is $2,048 \times 40 \times 16$ (the channel dimension is 2,048, the height dimension is 40, and the width dimension is 16), and average pooling is performed in the dimension in the first direction, thereby obtaining a pooled feature map of which a dimension is $2,048 \times 40 \times 1$.

[00107] In S222, the pooled feature map is split in the second direction to obtain the feature sequence.

[00108] In some embodiments, the pooled feature map is split in the dimension of the feature map in the second direction to obtain the feature sequence. The number of vectors obtained by splitting the pooled feature map may be determined based on a length of the feature map in the dimension in the second direction. For example, if the length of the feature map in the second direction is 60, the pooled feature map is split into 60 vectors. Each feature in the feature sequence corresponds to the same size information.

[00109] Based on S221 and S222, if the dimension of the pooled feature map is $2,048 \times 40 \times 1$, the pooled feature map is split in the dimension of the feature map in the second direction to obtain 40 2,048-dimensional vectors, of which each corresponds to a feature corresponding to $1/40$ of an image region in the second direction in the feature map. Accordingly, under the condition that the first direction is the width direction of the object sequence and the second direction is the height direction of the object sequence, the feature map is pooled in the first direction to obtain the pooled feature map, and the pooled feature map is split in the second direction, so that the feature sequence may retain more detail information of the image including the object sequence in the height direction.

[00110] In some embodiments, the feature map is pooled at first in the width dimension of the feature map, and then a splitting operation is performed on the pooled feature map corresponding to the feature map in the height dimension of the feature map, thereby splitting the feature map into the feature sequence. In this manner, feature extraction is performed on the image including the object sequence using the object sequence recognition network obtained by training based on an image similarity loss function and a

feature sequence alignment loss function to obtain the feature map, and the feature map is split according to the dimension information, so that the obtained feature sequence may retain more features in the height direction to make it easy to subsequently recognize the class of each object in the object sequence more accurately.

[00111] In some embodiments, the object sequence recognition network is configured to recognize the class of the object. The object sequence recognition network is obtained by training an object sequence recognition network to be trained. A training process of the object sequence recognition network to be trained may be implemented through the operations shown in FIG. 3. FIG. 3 is an implementation flowchart of a method for training an object sequence recognition network according to an embodiment of the application. The following descriptions will be made in combination with FIG. 3.

[00112] In S31, a sample image group is acquired.

[00113] In some embodiments, the sample image group may be image information collected by any electronic device with an image collection function. The sample image group includes at least two frames of sample images extracted from a video stream. Timing of each frame of sample image in each sample image group satisfies a preset timing condition. The position of the same sample object sequence in each frame of sample image in a sample image group satisfies a preset consistency condition. Each frame of sample image includes class labeling information of a sample object sequence.

[00114] Here, the timing of the sample image may be a timing position thereof in the video stream or collection time of the sample image. That the timing of the sample images satisfies the preset timing condition may refer to that a distance between timing positions of the sample images in the video stream is less than a preset threshold, or may refer to that a distance between collection time of the sample images is less than a preset threshold.

[00115] In some embodiments, each frame of sample image in the sample image group includes the same sample object sequence. Positions of the sample object sequence in multiple frames of sample images which are close in timing in the video stream may usually not change greatly. Therefore, multiple frames of images of which timing

satisfies the preset timing condition and in which positions of the same sample object sequence do not change greatly may be determined as multiple frames of similar images. The preset consistency condition refers to that a difference between the positions does not exceed a preset difference range. For example, continuous image frames in the video stream are detected to obtain a detection box of the object sequence in each frame of image, and whether positions of the detection box in multiple frames of continuous or discontinuous images change beyond the difference range is judged. Therefore, it may be determined that there are relatively high correlations and similarities between sample images in each sample image group, and furthermore, the accuracy of the object recognition network obtained by training based on the sample image group in an object sequence recognition task may be improved.

[00116] In some embodiments, the sample image group may be image information obtained by preprocessing. For example, each sample image in the sample image group is the same in image size and/or image pixel value.

[00117] In some embodiments, positions of the sample object sequence in pictures of the sample images in the sample image group are the same or similarities are greater than a preset threshold, and timing of the images in the sample image group satisfies the preset timing condition. Alternatively, regions of a detection box corresponding to the sample object sequence in the sample images are the same or similarities are greater than a preset threshold, and timing of the images in the image sample group satisfies the preset timing condition.

[00118] In some possible implementation modes, the sample image group may be obtained from a first sample video stream according to position information of the sample object sequence and timing information of the sample images. That is, S31 may be implemented through the following S311 to S313 (not shown in the figure).

[00119] In S311, a sample video stream including the sample object sequence is acquired.

[00120] In some embodiments, video collection may be performed at first on a scene with a sample object by a device with a video collection function to obtain a sample

video stream. Then, a class of the sample object sequence in each sample image in the sample video stream is labeled to obtain the sample video stream. The sample video stream may be a group of videos or a random combination of multiple groups of videos.

[00121] In S312, sample object sequence detection is performed on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images.

[00122] In some embodiments, the sample object sequence in a picture of each sample image in the sample video stream may be detected by a trained detection model to determine a detection box corresponding to the sample object sequence, thereby determining the sample position of the sample object sequence in each sample image based on position information of the detection box in each sample image. The sample position of the sample object sequence in each sample image may be represented by a two-dimensional coordinate.

[00123] In S313, at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images are determined to form the sample image group.

[00124] In some embodiments, at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in multiple frames of sample images are determined as a sample image group according to the sample positions of the sample object sequence in each sample image in the sample video stream and timing information of each sample image. Sample positions of the sample object sequence in corresponding images in each sample image group satisfy the preset consistency condition. In addition, there may be one or more than two sample image groups. Each sample image group may include two or more sample images.

[00125] In some embodiments, an image size and/or image pixel processing may be performed on sample images in any sample image group in multiple sample image groups. Then, data enhancement is performed on processed sample images in any sample

image group, e.g., horizontal flipping, random pixel disturbance addition, image resolution or brightness adjustment, clipping, image feature distortion or random aspect ratio fine adjustment, thereby obtaining multiple frames of images related to picture contents of each sample image in the sample image group. Meanwhile, the multiple frames of images may be combined with the sample images, thereby generating the sample image group. As such, the richness of sample image group data may be improved.

[00126] In some possible implementation modes, image parameter adjustment and data enhancement are sequentially performed on sample images in a sample image group, thereby obtaining a sample image group. That is, an image parameter of each sample image in the sample image group is preprocessed at first according to a preset image parameter to obtain an intermediate sample image group. Then, data enhancement is performed on each intermediate sample image in the intermediate sample image group to obtain the sample image group. An implementation process is similar to preprocessing an acquired video stream to obtain an image including an object sequence. Therefore, the richness of sample image group data may be improved, meanwhile, the overall robustness of the object sequence recognition network to be trained may be improved, and furthermore, the accuracy of recognizing each object in the object sequence in the picture of the image may be improved.

[00127] In S32, the sample image group is input to an object sequence recognition network to be trained, and feature extraction is performed to obtain sample feature sequences.

[00128] In some embodiments, feature extraction is performed on each sample image in the sample image group using a convolutional network obtained by finely adjusting a structure of a ResNet, thereby obtaining the sample feature sequence of each sample image.

[00129] In some possible implementation modes, feature extraction is performed at first on each sample image in the sample image group, and then a splitting operation is performed on a feature map, thereby obtaining the sample feature sequence. That is, S32 may be implemented through S321 and S322 (not shown in the figure).

[00130] In S321, feature extraction is performed on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group.

[00131] In some embodiments, the convolutional subnetwork in the object sequence recognition network to be trained may be the convolutional network obtained by finely adjusting the network structure of the ResNet. For example, a high-layer feature in each sample image in the sample image group may be extracted using the convolutional subnetwork in the object sequence recognition network to be trained, thereby obtaining the sample feature map of each sample image in the sample image group.

[00132] In some possible implementation modes, feature extraction may be performed on each sample image in the sample image group, thereby obtaining a feature map of which a width is changed and a height is kept unchanged. That is, S321 may be implemented through the following process.

[00133] First, each sample image in the sample image group is down-sampled using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature.

[00134] The first direction is different from a sequencing direction of sample objects in the sample object sequence.

[00135] Then, a feature in a length dimension of each sample image in the sample image group in a second direction is extracted based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature.

[00136] Finally, the sample feature map of each sample image in the sample image group is obtained based on the first-dimensional sample feature and the second-dimensional sample feature.

[00137] In some embodiments, the abovementioned implementation process is similar to that of S211 to S213 in the abovementioned embodiment. Under the condition that the

first direction is a width direction of the sample object sequence and the second direction is a height direction of the sample object sequence, it is set at first that width strides 2 in last strides of convolutional layers 3 and 4 in the convolutional subnetwork are kept unchanged and height strides are changed from 2 to 1, to obtain a first-dimensional sample feature and second-dimensional sample feature corresponding to the sample image. Then, the first-dimensional sample feature may be combined with the second-dimensional sample feature to obtain a feature map of each sample image in the sample image group. As such, feature information of each sample image in the height dimension may be maximally retained.

[00138] In S322, the sample feature map of each sample image in the sample image group is split to obtain the sample feature sequence of each sample image in the sample image group.

[00139] Here, an implementation process of S322 is similar to that of S202. That is, the sample feature map is processed differently based on the height dimension and the width dimension to obtain the sample feature sequence.

[00140] In some possible implementation modes, the sample feature map of each sample image is pooled in a dimension in the first direction to obtain a pooled sample feature map, and the obtained pooled sample feature map is split in a dimension in the second direction to obtain the sample feature sequence of each sample image. That is, S322 may be implemented through the following process.

[00141] First, the sample feature map of each sample image in the sample image group is pooled in the first direction to obtain a pooled sample feature map of each sample image in the sample image group.

[00142] Then, the pooled sample feature map of each sample image in the sample image group is split in the second direction to obtain the sample feature sequence of each sample image in the sample image group.

[00143] Here, the abovementioned implementation process is similar to that of S221 and S222. That is, the sample feature map of each sample image is split in the height dimension of the sample feature map to obtain the feature sequence of each sample

image. Accordingly, the sample feature map is split in the height direction after being pooled in the width direction, so that the sample feature sequence may include more detail information of each sample image in the height direction.

[00144] In some embodiments, feature extraction is performed on each sample image in the sample image group using the object sequence recognition network to be trained to obtain the sample feature map, and the sample feature map is split according to dimension information, so that the obtained sample feature sequence may retain more features in the height direction to make it easy to subsequently recognize a class of the sample object in the sample feature sequence more accurately.

[00145] In S33, class prediction is performed on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group.

[00146] In some embodiments, classes of the sample objects corresponding to sample features in the sample feature sequence of each sample image in the sample image group may be predicted using a classifier in the object sequence recognition network to be trained, thereby obtaining a predicted probability of the sample object corresponding to each sample feature.

[00147] In some embodiments, the sample feature sequence is input to the classifier of the object sequence recognition network to be trained, and class prediction is performed to obtain a sample classification result of each sample feature sequence.

[00148] In S34, a first loss and a second loss set are determined based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group.

[00149] In some embodiments, the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images. The similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images. A second loss in the second loss set is configured to represent a difference

between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence.

[00150] In some embodiments, the second loss for supervising a classification result of each sample object in the sample object sequence in each sample image may be determined according to a classification result of each sample object sequence output by the classifier in the object sequence recognition network to be trained and truth value information of a class of each sample object sequence to obtain the second loss set. The number of the second losses in the second loss set is the same as that of the sample images in the sample image group. In addition, the second loss set may be a CTC loss set.

[00151] In some embodiments, a CTC loss is adopted as the second loss, and a pair loss is adopted as the first loss. For each sample image in the sample image group, the second loss of the sample image is obtained taking the classification result of the sample feature sequence of the sample image output by the classifier and a truth value label of the class of the sample object sequence in the sample image as an input of the CTC loss to predict the class of each sample object in the sample feature sequence of the sample image. As such, the second loss set may be obtained based on the group of sample images. Meanwhile, the first loss for supervising similarities between multiple frames of different sample images in the sample image group is determined based on sample similarities between the multiple frames of different sample images in the sample image group and truth value similarities between different sample images in the sample image group. The first loss may be a pair loss.

[00152] In some embodiments, a pair loss is adopted as the first loss. For example, an implementation form of the pair loss may be selected from losses for measuring distribution differences, e.g., a Layer 2 (L2) loss, a cos loss, and a Kullback-Leibler divergence loss.

[00153] In S35, a network parameter of the object sequence recognition network to be trained is adjusted according to the first loss and the second loss set such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.

[00154] Here, similarities between different sample images in the sample image group may be compared with similarity truth values between different sample images to determine the first loss. The class representing each sample object in each sample object sequence in the predicted classes may be compared with class truth value information of each sample object of a sequence of sample objects belonging to the same class to determine the second loss set. The first loss and the second loss set are fused to adjust a weight value of the object sequence recognition network to be trained to converge losses of the classes of the sample objects output by the trained object sequence recognition network.

[00155] Through S31 to S36, the second loss set for supervising object sequences and the first loss for supervising similarities between different images in a group of sample images are introduced to the object sequence recognition network to be trained based on the image group, so that the feature extraction consistency of similar images may be improved, and furthermore, an overall class prediction effect of the network is improved.

[00156] In some possible implementation modes, the first loss and the second loss set are adjusted to obtain a total loss. Meanwhile, the network parameter of the object sequence recognition network to be trained is adjusted based on the total loss to obtain the object sequence recognition network. That is, S36 may be implemented through the following S361 and S362.

[00157] In S361, weighted fusion is performed on the first loss and the second loss set to obtain a total loss.

[00158] In some embodiments, the first loss and the second loss set are weighted using different weights respectively, and a first loss and second loss set which are obtained by weighted adjustment are fused to obtain the total loss.

[00159] In some possible implementation modes, preset adjustment parameters are set for the first loss and the second loss set to obtain the total loss. That is, S361 may be implemented through the following process.

[00160] In a first step, the first loss is adjusted using a first preset weight to obtain a third loss.

- [00161] In some embodiments, the first loss is adjusted using the first preset weight to obtain the third loss. The first weight may be a preset numerical value, or may be determined based on a parameter of the object sequence recognition network to be trained in the training process.
- [00162] In a second step, a class supervision weight is determined based on the number of the sample images in the sample image group.
- [00163] In some embodiments, multiple different sample images in the same image group correspond to the same class supervision weight. The class supervision weight is determined based on the number of the sample images in the sample image group. In such case, multiple class supervision weights may be the same numerical value or different numerical values, but a sum of the multiple class supervision weights is 1. For example, if the number of the sample images in the sample image group is n , the class supervision weight may be $1/n$.
- [00164] In some embodiments, if the number of the sample images in the sample image group is 2, the class supervision weight may be 0.5. Alternatively, if the number of the sample images in the sample image group is 3, the class supervision weight may be 0.33.
- [00165] In a third step, the second losses in the second loss set are fused based on the class supervision weight and a second preset weight to obtain a fourth loss.
- [00166] In some embodiments, there may be a preset relationship between the first preset weight and the second preset weight. For example, a ratio of the first preset weight to the second preset weight is fixed. Alternatively, a difference between the first preset weight and the second preset weight is fixed.
- [00167] In some embodiments, the second losses in the second loss set are adjusted based on the class supervision weight and the second preset weight to obtain the fourth loss. For example, the class supervision weight is multiplied by the second preset weight, each second loss in the second loss set is sequentially adjusted to further obtain an adjusted second loss set, and multiple losses in the adjusted second loss set are summed to obtain the fourth loss. Alternatively, the class supervision weight and the second preset weight are added, each second loss in the second loss set is sequentially adjusted to

further obtain an adjusted second loss set, and multiple losses in the adjusted second loss set are summed to obtain the fourth loss.

[00168] In some possible implementation modes, each second loss in the second loss set is adjusted through the class supervision weight, thereby obtaining the fourth loss. That is, the following implementation process may be adopted.

[00169] First, the class supervision weight is assigned to each second loss in the second loss set to obtain an updated loss set including at least two updated losses.

[00170] In some embodiments, the class supervision weight is assigned to each second loss in the second loss set to obtain the updated loss corresponding to each second loss. Furthermore, the updated loss set is obtained based on the updated loss corresponding to each second loss. There is a mapping relationship between each updated loss in the updated loss set and each second loss in the first loss set.

[00171] Then, the updated losses in the updated loss set are fused to obtain a fused loss.

[00172] In some embodiments, each updated loss in the updated loss set may be summed to obtain the fused loss.

[00173] Finally, the fused loss is adjusted using the second preset weight to obtain the fourth loss.

[00174] In some embodiments, when the fused loss is adjusted using the second preset weight, the second preset weight may be multiplied by the fused loss so as to obtain the fourth loss, or the second preset weight may be divided by the fused loss so as to obtain the fourth loss. The second preset weight may be a preset numerical value, or may be determined based on a parameter of the object sequence recognition network to be trained in the training process.

[00175] Here, the second loss set is adjusted sequentially through the class supervision weight associated with the number of the sample images in the sample image group and the first preset weight, thereby obtaining the fourth loss. As such, the second loss set for supervising classes of sample objects in a group of sample images may have relatively

high performance in the training process, and meanwhile, the network parameter of the object sequence recognition network to be trained may further be optimized.

[00176] In a fourth step, the total loss is determined based on the third loss and the fourth loss.

[00177] In some embodiments, the total loss is determined based on the third loss and the fourth loss. The total loss may be determined by adding the third loss and the fourth loss.

[00178] In S362, the network parameter of the object sequence recognition network to be trained is adjusted according to the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.

[00179] In some embodiments, the network parameter of the object sequence recognition network to be trained is adjusted using the total loss obtained by fusing the third loss and the fourth loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition. In this manner, the object sequence recognition network to be trained may be trained to improve the prediction effect of the whole network, so that an object sequence recognition network with relatively high performance may be obtained.

[00180] The method for training an object sequence recognition network will be described below in combination with a specific embodiment. For example, an application scene is a game place, and an object (for example, a token) in the game place is recognized. However, it is to be noted that the specific embodiment is only for describing the embodiments of the application better and not intended to form improper limits to the embodiments of the application.

[00181] A sequence recognition algorithm for an image is applied extensively to scene text recognition, license plate recognition and other scenes. In the related art, the algorithm mainly includes extracting an image feature using a convolutional neural network, performing classification prediction on each slice feature, performing duplicate elimination in combination with a CTC loss function and supervising a predicted output,

and is applicable to text recognition and license plate recognition tasks.

[00182] However, for the recognition of a token sequence in the game place, the stacked token sequence usually has a relatively great sequence length, and a requirement on the accuracy of predicting a face value and type of each token is relatively high, so an effect of performing sequence recognition on stacked tokens based on a DL method is not so good.

[00183] Based on this, an embodiment of the application provides an object sequence recognition method. A pair loss based on a feature similarity of paired images is added based on CTC-loss-based token recognition, so that the feature extraction consistency of similar images may be improved, and furthermore, each object in an object sequence may be recognized accurately.

[00184] FIG. 4 is a structure diagram of an object sequence recognition network according to an embodiment of the application. The following descriptions will be made in combination with FIG. 4. A framework of the object sequence recognition network includes a video frame group construction module 401, a feature extraction module 402, and a loss module.

[00185] The video frame group construction module 401 is configured to construct a corresponding video frame group for each video frame in training video stream data to obtain a sample video frame group.

[00186] Video stream data is usually taken as an input in a game place. However, an input for token recognition is usually an image region corresponding to a token detection box of a target detection model. In continuous video stream data, a token sequence video frame group including the same token information may be obtained through a certain screening condition, for example, detection box coordinates of a sample object sequence in continuous video frames are the same, based on timing information and detection box information corresponding to the sample object sequence, namely video frames of each group have the same label. Any two video frames in each group of video frames may form a video frame group to facilitate subsequent model training. In addition, more than two video frames may be selected from each group of video frames to form a

combination for training.

[00187] In addition, each video frame in the video frame group is further preprocessed, including adjusting a size of an image according to an aspect ratio, normalizing pixel values of the image, and other operations. The operation of adjusting the size of the image according to the aspect ratio refers to adjusting widths of multiple video frames to be the same. As such, great deformations of the multiple video frames generated if aspect ratios of the multiple video frames are not adjusted to be the same because tokens in the input video frames are different in number and the aspect ratios of images are greatly different may be reduced. For example, for an image of which an image height is less than a maximum height, other positions of which heights are less than the maximum height are filled with average gray pixel values (127, 127, 127). In order to enrich a sample image set, a data enhancement operation may further be performed on processed video frames, e.g., horizontal flipping, random pixel disturbance addition, image resolution or brightness adjustment, clipping, image feature distortion or random aspect ratio fine adjustment. As such, the overall robustness of the network to be trained may be improved.

[00188] The feature extraction module 402 performs feature extraction on video frames in a processed video frame group to obtain feature sequences 4031 and 4032.

[00189] High-layer features of the input video frames are extracted at first using a convolutional neural network part in the object sequence recognition network to be trained. The convolutional neural network part is obtained by fine adjustment based on a network structure of a ResNet. For example, last strides of convolutional layers 3 and 4 in the network structure of the ResNet are changed from (2, 2) to (1, 2). As such, an obtained feature map is not down-sampled in a height dimension, and is down-sampled in a width dimension to be halved, namely a feature map of each video frame in the video frame group is obtained. Therefore, feature information in the height dimension may be maximally retained. Then, a splitting operation is performed on the feature map of each video frame in the video frame group, namely the feature map extracted by the convolutional neural network is split into a plurality of feature sequences to facilitate subsequent calculation of a classifier and a loss function. When the feature map is split,

average pooling is performed in a width direction of the feature map, and no changes are made in a height direction and a channel dimension. For example, a size of the feature map is 2,048*40*8 (the channel dimension is 2,048, the height dimension is 40, and the width dimension is 8), a 2,048*40*1 feature map is obtained by average pooling in the width direction, and the feature map is split in the height dimension to obtain 40 2,048-dimensional vectors, of which each corresponds to a feature corresponding to 1/40 of a region in the height direction in the original map.

[00190] If a sample image includes multiple tokens, as shown in FIG. 5 which is a schematic diagram of an application scene of an object sequence recognition network according to an embodiment of the application, the feature sequence is obtained by division according to a height dimension of the image 501. A feature sequence includes a feature of less than or equal to one token.

[00191] Then, a class of each object in an object sequence of each video frame in the video frame group is predicted using an n-classifier, thereby obtaining a predicted probability of each feature sequence. Here, n is the total number of token classes.

[00192] Meanwhile, similarities, i.e., feature similarities 404, between different video frames in the video frame group may further be determined.

[00193] For the feature sequence obtained by the convolutional network, the loss module determines feature similarities between different video frames in the video frame group using a pair loss 406 and supervises the network for an optimization purpose of improving the similarities. For predicted probabilities of all feature sequence classes, a prediction result of the object sequence of each video frame in the video frame group is supervised using a CTC loss 405 and a CTC loss 407 respectively.

[00194] The pair loss 406, the CTC loss 405 and the CTC loss 407 are fused to obtain a total loss 408. For example, the total loss corresponding to the pair loss L_{pair} 406, the CTC loss L_{ctc1} 405 and the CTC loss L_{ctc2} 407 (for example, the video frame group includes two video frames) is $L = \alpha(0.5L_{ctc1} + 0.5L_{ctc2}) + \beta L_{pair}$. Meanwhile, the pair loss 406 may be selected from losses for measuring distribution differences. Values of α and

β may be set as $\alpha:\beta=1:10$ based on a practical application.

[00195] Finally, back propagation is performed according to the classification result of the feature sequence and calculation results of the loss functions to update a network parameter weight. In a test stage, the classification result of the feature sequence is processed according to a post-processing rule of the CTC loss function to obtain a predicted token sequence result, including a length of the token sequence and a class corresponding to each token.

[00196] In the embodiment of the application, without introducing any additional parameter or modifying a network structure, the prediction result of the sequence length may be improved, and meanwhile, the accuracy of recognizing the class of the object may be improved to finally improve the overall recognition result, particularly in a scene including stacked tokens.

[00197] Based on the abovementioned embodiments, an embodiment of the application provides an object sequence recognition apparatus. FIG. 6A is a structure composition diagram of an object sequence recognition apparatus according to an embodiment of the application. As shown in FIG. 6A, the object sequence recognition apparatus 600 includes a first acquisition module 601, a first extraction module 602, and a first determination module 603.

[00198] The first acquisition module 601 is configured to acquire an image including an object sequence.

[00199] The first extraction module 602 is configured to perform feature extraction on the image including the object sequence using an object sequence recognition network to obtain a feature sequence. Supervision information in a training process of the object sequence recognition network at least include first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image. Each sample image group includes at least two frames of sample images extracted from the same video stream. Timing of each frame of sample image in each sample image group satisfies a preset timing condition. The position of the same sample object

sequence in each frame of sample image in a sample image group satisfies a preset consistency condition.

[00200] The first determination module 603 is configured to determine a class of each object in the object sequence based on the feature sequence.

[00201] In some embodiments, the first extraction module 602 includes a first feature extraction submodule and a first splitting submodule. The first feature extraction submodule is configured to perform feature extraction on the image including the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map. The first splitting submodule is configured to split the feature map to obtain the feature sequence.

[00202] In some embodiments, the first feature extraction submodule includes a first down-sampling subunit, a first feature extraction subunit, and a first feature map determination subunit. The first down-sampling subunit is configured to down-sample the image including the object sequence using the convolutional subnetwork in a length dimension of the image including the object sequence in a first direction to obtain a first-dimensional feature, the first direction being different from an arrangement direction of the objects in the object sequence. The first feature extraction subunit is configured to extract a feature in a length dimension of the image including the object sequence in a second direction based on a length of the image including the object sequence in the second direction to obtain a second-dimensional feature. The first feature map determination subunit is configured to obtain the feature map based on the first-dimensional feature and the second-dimensional feature.

[00203] In some embodiments, the first splitting submodule includes a first pooling subunit and a first splitting subunit. The first pooling subunit is configured to pool the feature map in the first direction to obtain a pooled feature map. The first splitting subunit is configured to split the pooled feature map in the second direction to obtain the feature sequence.

[00204] An embodiment of the application also provides an apparatus for training an object sequence recognition network. FIG. 6B is a structure composition diagram of an

apparatus for training an object sequence recognition network according to an embodiment of the application. As shown in FIG. 6B, the apparatus 610 for training an object sequence recognition network includes a second acquisition module 611, a second extraction module 612, a second prediction module 613, a second determination module 614, and a first adjustment module 615.

[00205] The second acquisition module 611 is configured to acquire a sample image group. The sample image group includes at least two frames of sample images extracted from the same video stream. Timing of each frame of sample image in each sample image group satisfies a preset timing condition. A position of the same sample object sequence in each frame of sample image in a sample image group satisfies a preset consistency condition. Each frame of sample image includes class labeling information of a sample object sequence.

[00206] The second extraction module 612 is configured to input the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain sample feature sequences.

[00207] The second prediction module 613 is configured to perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group.

[00208] The second determination module 614 is configured to determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group. The first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images. The similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images. A second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence.

[00209] The first adjustment module 615 is configured to adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.

[00210] In some embodiments, the second acquisition module 611 includes a second acquisition submodule, a second detection submodule, and a second forming submodule. The second acquisition submodule is configured to acquire a sample video stream including the sample object sequence. The second detection submodule is configured to perform sample object sequence detection on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images. The second forming submodule is configured to determine at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images to form the sample image group.

[00211] In some embodiments, the second extraction module 612 includes a second feature extraction submodule and a second splitting submodule. The second feature extraction submodule is configured to perform feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group. The second splitting submodule is configured to split the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the sample image group.

[00212] In some embodiments, the second feature extraction submodule includes a second down-sampling subunit, a second feature extraction subunit, and a second feature map determination subunit. The second down-sampling subunit is configured to down-sample each sample image in the sample image group using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature, the first direction being different from an arrangement direction of sample objects in the sample object sequence. The second feature extraction

subunit is configured to extract a feature in a length dimension of each sample image in the sample image group in a second direction based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature. The second feature map determination subunit is configured to obtain the sample feature map of each sample image in the sample image group based on the first-dimensional sample feature and the second-dimensional sample feature.

[00213] In some embodiments, the second splitting submodule includes a second pooling subunit and a second splitting subunit. The second pooling subunit is configured to pool the sample feature map of each sample image in the sample image group in the first direction to obtain a pooled sample feature map of each sample image in the sample image group. The second splitting subunit is configured to split the pooled sample feature map of each sample image in the sample image group in the second direction to obtain the sample feature sequence of each sample image in the sample image group.

[00214] In some embodiments, the first adjustment module 615 includes a fusion submodule and an adjustment submodule. The fusion submodule is configured to perform weighted fusion on the first loss and the second loss set to obtain a total loss. The adjustment submodule is configured to adjust the network parameter of the object sequence recognition network to be trained according to the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.

[00215] In some embodiments, the fusion submodule includes a first adjustment unit, a weight determination unit, a fusion unit, and a determination unit. The first adjustment unit is configured to adjust the first loss using a first preset weight to obtain a third loss. The weight determination unit is configured to determine a class supervision weight based on the number of the sample images in the sample image group, multiple different sample images in the same sample image group corresponding to the same class supervision weight. The fusion unit is configured to fuse the second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss. The determination unit is configured to determine the total loss based on the third loss and the fourth loss.

[00216] In some embodiments, the fusion unit includes an assignment subunit, a fusion subunit, and an adjustment subunit. The assignment subunit is configured to assign the class supervision weight to each second loss in the second loss set to obtain an updated loss set including at least two updated losses. The fusion subunit is configured to fuse the updated losses in the updated loss set to obtain a fused loss. The adjustment subunit is configured to adjust the fused loss using the second preset weight to obtain the fourth loss.

[00217] It is to be pointed out that descriptions about the above apparatus embodiment are similar to those about the method embodiment, and beneficial effects similar to those of the method embodiment are achieved. Technical details undisclosed in the apparatus embodiment of the application may be understood with reference to the descriptions about the method embodiment of the application.

[00218] It is to be noted that, in the embodiments of the application, the body and hand correlation method may also be stored in a computer-readable storage medium when being implemented in form of a software function module and sold or used as an independent product. Based on such an understanding, the technical solutions of the embodiments of the application substantially or parts making contributions to the related art may be embodied in form of a software product. The computer software product is stored in a storage medium, including a plurality of instructions configured to enable an electronic device (which may be a smart phone with a camera, a tablet computer, etc.) to execute all or part of the method in each embodiment of the application. The storage medium includes various media capable of storing program codes such as a U disk, a mobile hard disk, a Read Only Memory (ROM), a magnetic disk, or an optical disk. Therefore, the embodiments of the application are not limited to any specific hardware and software combination.

[00219] Based on the same technical concept, an embodiment of the application provides a computer device, which is configured to implement the object sequence recognition method and the method for training an object sequence recognition network in the method embodiments. FIG. 5 is a composition structure diagram of a computer device according to an embodiment of the application. As shown in FIG. 7, the computer

device 700 includes a processor 701, at least one communication bus, a communication interface 702, at least one external communication interface, and a memory 703. The communication interface 702 is configured to implement connections and communications between these components. The communication interface 702 may include a display screen. The external communication interface may include a standard wired interface and wireless interface. The processor 701 is configured to execute an object recognition program and object recognition network training program in the memory to implement the object sequence recognition method and the method for training an object sequence recognition network in the abovementioned embodiments.

[00220] Correspondingly, an embodiment of the application provides a computer-readable storage medium having stored therein a computer program which is executed by a processor to implement any object recognition method and method for training an object sequence recognition network in the abovementioned embodiments.

[00221] Correspondingly, an embodiment of the application also provides a chip, which includes a programmable logic circuit and/or a program instruction and is configured to, when running, implement any object recognition method and method for training an object sequence recognition network in the abovementioned embodiments.

[00222] Correspondingly, an embodiment of the application also provides a computer program product which, when being executed by a processor of an electronic device, is configured to implement any object recognition method and method for training an object sequence recognition network in the abovementioned embodiments.

[00223] The above descriptions about the embodiments of the object recognition apparatus, the computer device and the storage medium are similar to the descriptions about the method embodiments, and technical descriptions and beneficial effects are similar to those of the corresponding method embodiments. Due to the space limitation, references can be made to the records in the method embodiments, and elaborations are omitted herein. Technical details undisclosed in the embodiments of the object recognition apparatus, computer device and storage medium of the application may be understood with reference to the descriptions about the method embodiments of the application.

[00224] It is to be understood that "one embodiment" and "an embodiment" mentioned in the whole specification mean that specific features, structures or characteristics related to the embodiment is included in at least one embodiment of the application. Therefore, "in one embodiment" or "in an embodiment" mentioned throughout the specification does not always refer to the same embodiment. In addition, these specific features, structures or characteristics may be combined in one or more embodiments freely as appropriate. It is to be understood that, in each embodiment of the application, the magnitude of the sequence number of each process does not mean an execution sequence and the execution sequence of each process should be determined by its function and an internal logic and should not form any limit to an implementation process of the embodiments of the application. The sequence numbers of the embodiments of the application are adopted not to represent superiority-inferiority of the embodiments but only for description. It is to be noted that terms "include" and "contain" or any other variant thereof is intended to cover nonexclusive inclusions herein, so that a process, method, object, or device including a series of elements not only includes those elements but also includes other elements which are not clearly listed or further includes elements intrinsic to the process, the method, the object, or the device. Under the condition of no more limitations, an element defined by the statement "including a/an....." does not exclude existence of the same other elements in a process, method, object, or device including the element.

[00225] In some embodiments provided by the application, it is to be understood that the disclosed device and method may be implemented in another manner. The device embodiment described above is only schematic, and for example, division of the units is only logic function division, and other division manners may be adopted during practical implementation. For example, multiple units or components may be combined or integrated into another system, or some characteristics may be neglected or not executed. In addition, coupling or direct coupling or communication connection between each displayed or discussed component may be indirect coupling or communication connection, implemented through some interfaces, of the device or the units, and may be electrical and mechanical or adopt other forms.

[00226] The units described as separate parts may or may not be physically separated,

and parts displayed as units may or may not be physical units, and namely may be located in the same place, or may also be distributed to multiple network units. Part of all of the units may be selected according to a practical requirement to achieve the purposes of the solutions of the embodiments.

[00227] In addition, each function unit in each embodiment of the application may be integrated into a processing unit, each unit may also serve as an independent unit and two or more than two units may also be integrated into a unit. The integrated unit may be implemented in a hardware form and may also be implemented in form of hardware and software function unit. Those of ordinary skill in the art should know that all or part of the steps of the method embodiment may be implemented by related hardware instructed through a program, the program may be stored in a computer-readable storage medium, and the program is executed to execute the steps of the method embodiment. The storage medium includes various media capable of storing program codes such as a mobile storage device, a ROM, a magnetic disk, or an optical disc.

[00228] Or, the integrated unit of the application may also be stored in a computer-readable storage medium when being implemented in form of a software function module and sold or used as an independent product. Based on such an understanding, the technical solutions of the embodiments of the application substantially or parts making contributions to the conventional art may be embodied in form of software product, and the computer software product is stored in a storage medium, including a plurality of instructions configured to enable a computer device (which may be a personal computer, a server, a network device or the like) to execute all or part of the method in each embodiment of the application. The storage medium includes various media capable of storing program codes such as a mobile hard disk, a ROM, a magnetic disk, or an optical disc. The above is only the specific implementation mode of the application and not intended to limit the scope of protection of the application. Any variations or replacements apparent to those skilled in the art within the technical scope disclosed by the application shall fall within the scope of protection of the application. Therefore, the scope of protection of the application shall be subject to the scope of protection of the claims.

CLAIMS

1. An object sequence recognition method, comprising:

acquiring an image comprising an object sequence;

performing feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition; and

determining a class of each object in the object sequence based on the feature sequence.

2. The method of claim 1, wherein the performing feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence comprises:

performing feature extraction on the image comprising the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map; and

splitting the feature map to obtain the feature sequence.

3. The method of claim 2, wherein the performing feature extraction on the image comprising

the object sequence using a convolutional subnetwork in the object sequence recognition network to obtain a feature map comprises:

down-sampling the image comprising the object sequence using the convolutional subnetwork in a length dimension of the image comprising the object sequence in a first direction to obtain a first-dimensional feature, the first direction being different from an arrangement direction of objects in the object sequence;

extracting a feature in a length dimension of the image comprising the object sequence in a second direction based on a length of the image comprising the object sequence in the second direction to obtain a second-dimensional feature; and

obtaining the feature map based on the first-dimensional feature and the second-dimensional feature.

4. The method of claim 3, wherein the splitting the feature map to obtain the feature sequence comprises:

pooling the feature map in the first direction to obtain a pooled feature map; and

splitting the pooled feature map in the second direction to obtain the feature sequence.

5. A method for training an object sequence recognition network, comprising:

acquiring a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence;

inputting the sample image group to an object sequence recognition network to be trained, and performing feature extraction to obtain sample feature sequences;

performing class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determining a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjusting a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

6. The method of claim 5, wherein the acquiring a sample image group comprises:

acquiring a sample video stream comprising the sample object sequence;

performing sample object sequence detection on multiple frames of sample images in the sample video stream to obtain a sample position of the sample object sequence in each frame of sample image in the multiple frames of sample images; and

determining at least two frames of sample images which satisfy the preset timing condition and in which the sample positions of the sample object sequence satisfy the preset consistency condition in the multiple frames of sample images to form the sample image group.

7. The method of claim 5 or 6, wherein the inputting the sample image group to an object sequence recognition network to be trained and performing feature extraction to obtain sample feature sequences comprises:

performing feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group; and

splitting the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the sample image group.

8. The method of claim 7, wherein the performing feature extraction on each sample image in the sample image group using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image in the sample image group comprises:

down-sampling each sample image in the sample image group using the convolutional subnetwork in a length dimension of each sample image in a first direction to obtain a first-dimensional sample feature, the first direction being different from an arrangement direction of sample objects in the sample object sequence;

extracting a feature in a length dimension of each sample image in the sample image group in a second direction based on a length of each sample image in the sample image group in the second direction to obtain a second-dimensional sample feature; and

obtaining the sample feature map of each sample image in the sample image group based on the first-dimensional sample feature and the second-dimensional sample feature.

9. The method of claim 8, wherein the splitting the sample feature map of each sample image in the sample image group to obtain the sample feature sequence of each sample image in the

sample image group comprises:

pooling the sample feature map of each sample image in the sample image group in the first direction to obtain a pooled sample feature map of each sample image in the sample image group; and

splitting the pooled sample feature map of each sample image in the sample image group in the second direction to obtain the sample feature sequence of each sample image in the sample image group.

10. The method of any one of claims 5-9, wherein the adjusting a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition comprises:

performing weighted fusion on the first loss and the second loss set to obtain a total loss; and

adjusting the network parameter of the object sequence recognition network to be trained according to the total loss to make the loss of the classification result output by the adjusted object sequence recognition network satisfy the convergence condition.

11. The method of claim 10, wherein the performing weighted fusion on the first loss and the second loss set to obtain a total loss comprises:

adjusting the first loss using a first preset weight to obtain a third loss;

determining a class supervision weight based on the number of the sample images in the sample image group, multiple different sample images in the same sample image group corresponding to the same class supervision weight;

fusing second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss; and

determining the total loss based on the third loss and the fourth loss.

12. The method of claim 11, wherein the fusing second losses in the second loss set based on the class supervision weight and a second preset weight to obtain a fourth loss comprises:

assigning the class supervision weight to each second loss in the second loss set to obtain an updated loss set comprising at least two updated losses;

fusing the updated losses in the updated loss set to obtain a fused loss; and

adjusting the fused loss using the second preset weight to obtain the fourth loss.

13. An object sequence recognition apparatus, comprising:

a first acquisition module, configured to acquire an image comprising an object sequence;

a first extraction module, configured to perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition; and

a first determination module, configured to determine a class of each object in the object sequence based on the feature sequence.

14. An apparatus for training an object sequence recognition network, comprising:

a second acquisition module, configured to acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence;

a second extraction module, configured to input the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain sample feature sequences;

a second prediction module, configured to perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

a second determination module, configured to determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

a first adjustment module, configured to adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

15. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition; and

determine a class of each object in the object sequence based on the feature sequence.

16. A computer device, comprising a memory and a processor, wherein a computer-executable instruction is stored in the memory; and when executing the computer-executable instruction in the memory, the processor is configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

17. A computer storage medium, in which a computer-executable instruction is stored, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a

preset consistency condition; and

determine a class of each object in the object sequence based on the feature sequence.

18. A computer storage medium, in which a computer-executable instruction is stored, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling information of a sample object sequence;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained

according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.

19. A computer program, comprising computer instructions executable by an electronic device, wherein when executed by a processor in the electronic device, the computer instructions are configured to:

acquire an image comprising an object sequence;

perform feature extraction on the image comprising the object sequence using an object sequence recognition network to obtain a feature sequence, wherein supervision information in a training process of the object sequence recognition network at least comprises first supervision information for a similarity between at least two frames of sample images in a sample image group and second supervision information for a class of a sample object sequence in each sample image, each sample image group comprises at least two frames of sample images extracted from a video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, and positions of a sample object sequence in the frames of sample image in a sample image group satisfies a preset consistency condition; and

determine a class of each object in the object sequence based on the feature sequence.

20. A computer program, comprising computer instructions executable by an electronic device, wherein when executed by a processor in the electronic device, the computer instructions are configured to:

acquire a sample image group, wherein the sample image group comprises at least two frames of sample images extracted from the same video stream, timing of each frame of sample image in each sample image group satisfies a preset timing condition, positions of a sample object sequence in the frames of sample image in a sample image group satisfy a preset consistency condition, and each frame of sample image comprises class labeling

information of a sample object sequence;

input the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain sample feature sequences;

perform class prediction on sample objects in the sample feature sequences to obtain a predicted class of each sample object in the sample object sequence in each sample image in the sample image group;

determine a first loss and a second loss set based on the predicted class of each sample object in the sample object sequence in each sample image in the sample image group, wherein the first loss is negatively correlated with similarities between multiple frames of different sample images in the sample images, the similarities between the multiple frames of different sample images are determined based on sample feature sequences of the multiple frames of different sample images and/or predicted classes of sample object sequences in the multiple frames of different sample images, and a second loss in the second loss set is configured to represent a difference between the class labeling information of the sample object sequence in each frame of sample image and the predicted class of each sample object in the sample object sequence; and

adjust a network parameter of the object sequence recognition network to be trained according to the first loss and the second loss set to make a loss of a classification result output by an adjusted object sequence recognition network satisfy a convergence condition.