

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 of sequential objects. The neural network may be obtained by training taking classes of objects in sequential objects as supervision information.

[0003] In some scenes, object sequences are relatively long, and requirements on the accuracy of recognizing these objects are relatively high, so it is unlikely to achieve satisfactory sequence recognition effects by a sequence recognition method in the related art.

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] A first image including an object sequence is acquired.

[0008] The first image is input to an object sequence recognition network, and feature extraction is performed to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least including first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image.

[0009] A class of each object in the object sequence is determined based on the feature sequence.

[0010] In some embodiments, the operation that the first image is input to an object sequence recognition network and feature extraction is performed to obtain a feature sequence may include the following operations. Feature extraction is performed on the first image 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. Accordingly, the feature map is split according to dimension information, so that the obtained feature sequence may retain more features in a height direction to make it easy to subsequently recognize a class of the object sequence in the feature sequence more accurately.

[0011] In some embodiments, the operation that feature extraction is performed on the first image using a convolutional subnetwork in the object sequence recognition network to obtain a feature map may include the following operations. The first image is down-sampled using the convolutional subnetwork in a length dimension of the first image 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 first image in a second direction is extracted based on a length of the first image 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 first image in the dimension in the first 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 first direction after being pooled in the second direction, so that the feature sequence may include more detail information of the first image in the first direction.

[0013] In some embodiments, the operation that a class of each object in the object sequence is determined based on the feature sequence may include the following operations. A class corresponding to each feature in the feature sequence is predicted using a classifier of the object sequence recognition network. The class of each object in the object sequence is determined based on a prediction result of the class corresponding to each feature in the feature sequence. As such, a fixed-length feature sequence is converted to a variable feature sequence length.

[0014] 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, at least two frames of sample images in the sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image including a sample object sequence.

[0015] The sample images in the sample image group are input to an object sequence recognition network to be trained, and feature extraction is performed to obtain a sample feature sequence of each sample image in the sample image group. A class of a sample object sequence in each sample image is predicted based on the sample feature sequence of each sample image. A first loss for supervising the class of the sample object sequence in each sample image is determined based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set. A second loss for supervising a similarity between the at least two frames of sample images is determined based on the

sample feature sequences of the at least two frames of sample images in the sample image group. A network parameter of the object sequence recognition network to be trained is adjusted using the first loss set and the second loss such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition. Accordingly, the first loss set for supervising whole sequences and the second loss for supervising the similarity between a similarity between images in a group of sample images are introduced to a training process, so that the feature extraction consistency of similar images may be improved, and an overall class prediction effect of the network may be improved.

[0016] In some embodiments, the operation that a sample image group is acquired may include the following operations. A first sample image where a class of a sample object in a picture is labeled is acquired. At least one second sample image is determined based on a picture content of the first sample image. Data enhancement is performed on the at least one second sample image to obtain at least one third sample image. The sample image group is obtained based on the first sample image and the at least one third sample image. As such, paired images with similar picture contents are created through each frame of first sample image to make it easy to subsequently improve the feature extraction consistency of similar images.

[0017] In some embodiments, the operation that the sample images in the sample image group are input to an object sequence recognition network to be trained and feature extraction is performed to obtain a sample feature sequence of each sample object in the sample image group may include the following operations. Feature extraction is performed on each sample image using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image. The sample feature map of each sample image is split to obtain the sample feature sequence of each sample image. As such, the obtained sample feature sequence may retain more features in the first direction, and the training accuracy of the network may be improved.

[0018] In some embodiments, the operation that feature extraction is performed on each sample image using a convolutional subnetwork in the object sequence recognition

network to be trained to obtain a sample feature map of each sample image may include the following operations. Each sample image 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. Feature extraction is performed in a length dimension of each sample image in a second direction based on a length of each sample image in the second direction to obtain a second-dimensional sample feature. The sample feature map of each sample image is obtained based on the first-dimensional sample feature and the second-dimensional sample feature. As such, feature information in a dimension of each sample image in the first direction may be maximally retained.

[0019] In some embodiments, the operation that the sample feature map of each sample image is split to obtain the sample feature sequence of each sample image may include the following operations. The sample feature map is pooled in the first direction to obtain a pooled sample feature map. The pooled sample feature map is split in the second direction to obtain the sample feature sequence. Accordingly, the sample feature map is split in the dimension in the first direction after being pooled in the dimension in the second direction, so that the sample feature sequence may retain more detailed information of the sample image in the dimension in the first direction.

[0020] In some embodiments, the operation that a network parameter of the object sequence recognition network to be trained is adjusted using the first loss set and the second loss 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 set and the second loss to obtain a total loss. The network parameter of the object sequence recognition network to be trained is adjusted based on the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition. Accordingly, the first loss set and the second loss are fused to train the network, so that the robustness of a trained network is improved.

[0021] In some embodiments, the operation that weighted fusion is performed on the

first loss set and the second loss to obtain a total loss may include the following operations. A class supervision weight corresponding to the sample image group is determined based on the number of the sample images in the sample image group. The first losses in the first loss set of the sample image group are fused based on the class supervision weight and a first preset weight to obtain a third loss, The second loss is adjusted using 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 the feature extraction consistency of similar images may be improved, and the prediction effect of the whole network may be improved.

[0022] In some embodiments, the operation that the first losses in the first loss set of the sample image group are fused based on the class supervision weight and a first preset weight to obtain a third loss may include the following operations. The class supervision weight is assigned to each first loss in the first 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 first preset weight to obtain the third loss. Accordingly, 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.

[0023] 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.

[0024] The first acquisition module may be configured to acquire a first image including an object sequence.

[0025] The first extraction module may be configured to input the first image to an object sequence recognition network and perform feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least including first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least

two frames of sample images in each sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image.

[0026] The first determination module may be configured to determine a class of each object in the object sequence based on the feature sequence.

[0027] In some embodiments, the first extraction module may include a first extraction submodule and a first splitting submodule.

[0028] The first extraction submodule may be configured to perform feature extraction on the first image using a convolutional subnetwork in the object sequence recognition network to obtain a feature map.

[0029] The first splitting submodule may be configured to split the feature map to obtain the feature sequence.

[0030] In some embodiments, the first extraction submodule may include a first down-sampling unit, a first extraction unit, and a first determination unit.

[0031] The first down-sampling unit may be configured to down-sample the first image using the convolutional subnetwork in a length dimension of the first image 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.

[0032] The first extraction unit may be configured to extract a feature in a length dimension of the first image in a second direction based on a length of the first image in the second direction to obtain a second-dimensional feature.

[0033] The first determination unit may be configured to obtain the feature map based on the first-dimensional feature and the second-dimensional feature.

[0034] In some embodiments, the first splitting submodule may include a first pooling unit and a first splitting unit.

[0035] The first pooling unit may be configured to pool the feature map in the first

direction to obtain a pooled feature map.

[0036] The first splitting unit may be configured to split the pooled feature map in the second direction to obtain the feature sequence.

[0037] In some embodiments, the first determination module may include a first prediction submodule and a first determination submodule.

[0038] The first prediction submodule may be configured to predict a class corresponding to each feature in the feature sequence using a classifier of the object sequence recognition network.

[0039] The first determination submodule may be configured to determine the class of each object in the object sequence based on a prediction result of the class corresponding to each feature in the feature sequence.

[0040] 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 first prediction module, a second determination module, a third determination module, and a first adjustment module.

[0041] The second acquisition module may be configured to acquire a sample image group, at least two frames of sample images in the sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image including a sample object sequence.

[0042] The second extraction module may be configured to input the sample images in the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain a sample feature sequence of each sample image in the sample image group.

[0043] The first prediction module may be configured to predict a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image.

- [0044] The second determination module may be configured to determine a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set.
- [0045] The third determination module may be configured to determine a second loss for supervising a similarity between the at least two frames of sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group.
- [0046] The first adjustment module may be configured to adjust a network parameter of the object sequence recognition network to be trained using the first loss set and the second loss such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.
- [0047] In some embodiments, the second acquisition module may include a first acquisition submodule, a second determination submodule, a first enhancement submodule, and a third determination submodule.
- [0048] The first acquisition submodule may be configured to acquire a first sample image where a class of a sample object in a picture is labeled.
- [0049] The second determination submodule may be configured to determine at least one second sample image based on a picture content of the first sample image.
- [0050] The first enhancement submodule may be configured to perform data enhancement on the at least one second sample image to obtain at least one third sample image.
- [0051] The third determination submodule may be configured to obtain the sample image group based on the first sample image and the at least one third sample image.
- [0052] In some embodiments, the second extraction module may include a second extraction submodule and a second splitting submodule.
- [0053] The second extraction submodule may be configured to perform feature

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

[0054] The second splitting submodule may be configured to split the sample feature map of each sample image to obtain the sample feature sequence of each sample image.

[0055] In some embodiments, the second extraction submodule may include a second down-sampling unit, a second extraction unit, and a second determination unit.

[0056] The second down-sampling unit may be configured to down-sample each sample image 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.

[0057] The second extraction unit may be configured to perform feature extraction in a length dimension of each sample image in a second direction based on a length of each sample image in the second direction to obtain a second-dimensional sample feature.

[0058] The second determination unit may be configured to obtain the sample feature map of each sample image based on the first-dimensional sample feature and the second-dimensional sample feature.

[0059] In some embodiments, the second splitting submodule may include a second pooling unit and a second splitting unit.

[0060] The second pooling unit may be configured to pool the sample feature map in the first direction to obtain a pooled sample feature map.

[0061] The second splitting unit may be configured to split the pooled sample feature map in the second direction to obtain the sample feature sequence.

[0062] In some embodiments, the first adjustment module may include a first fusion submodule and a first adjustment submodule.

- [0063] The first fusion submodule may be configured to perform weighted fusion on the first loss set and the second loss to obtain a total loss.
- [0064] The first adjustment submodule may be configured to adjust the network parameter of the object sequence recognition network to be trained based on the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.
- [0065] In some embodiments, the first fusion submodule may include a third determination unit, a first fusion unit, a first adjustment unit, and a fourth determination unit.
- [0066] The third determination unit may be configured to determine a class supervision weight corresponding to the sample image group based on the number of the sample images in the sample image group.
- [0067] The first fusion unit may be configured to fuse the first losses in the first loss set of the sample image group based on the class supervision weight and a first preset weight to obtain a third loss.
- [0068] The first adjustment unit may be configured to adjust the second loss using a second preset weight to obtain a fourth loss.
- [0069] The fourth determination unit may be configured to determine the total loss based on the third loss and the fourth loss.
- [0070] In some embodiments, the first fusion unit may include a first assignment subunit, a first fusion subunit, and a first adjustment subunit.
- [0071] The first assignment subunit may be configured to assign the class supervision weight to each first loss in the first loss set to obtain an updated loss set including at least two updated losses.
- [0072] The first fusion subunit may be configured to fuse the updated losses in the updated loss set to obtain a fused loss.

[0073] The first adjustment subunit may be configured to adjust the fused loss using the first preset weight to obtain the third loss.

[0074] Correspondingly, 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.

[0075] 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.

[0076] According to the object sequence recognition method, network training method, apparatuses, device and storage medium provided in the embodiments of the application. First, feature extraction is performed on the first image using the object sequence recognition network adopting supervision information including the supervision on a similarity between a group of sample images and the supervision on classes of sample objects in the group of sample images to obtain the feature sequence, so that the feature extraction consistency of multiple frames of similar first images may be improved. Then, class prediction is performed on the object sequence in the feature sequence, so that a classification result of the object sequence in the obtained feature sequence is relatively accurate. Finally, the classification result of the object sequence in the feature sequence is further processed to determine the class of the object 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 may be improved.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0077] FIG. 1 is an implementation flowchart of an object sequence recognition method according to an embodiment of the application.
- [0078] FIG. 2A is another implementation flowchart of an object sequence recognition method according to an embodiment of the application.
- [0079] FIG. 2B is an implementation flowchart of a method for training an object sequence recognition network according to an embodiment of the application.
- [0080] FIG. 3 is a structure diagram of an object sequence recognition network according to an embodiment of the application.
- [0081] FIG. 4 is a schematic diagram of an application scene of an object sequence recognition network according to an embodiment of the application.
- [0082] FIG. 5A is a structure composition diagram of an object sequence recognition apparatus according to an embodiment of the application.
- [0083] FIG. 5B is a structure composition diagram of an apparatus for training an object sequence recognition network according to an embodiment of the application.
- [0084] FIG. 6 is a composition structure diagram of a computer device according to an embodiment of the application.

DETAILED DESCRIPTION

- [0085] 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.
- [0086] "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.

[0087] 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.

[0088] 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.

[0089] 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.

[0090] 1) Pair loss: Paired samples are used for loss calculation in many metric learning methods in deep learning. 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.

[0091] 2) Connectionist Temporal Classification (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.

[0092] 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.

[0093] 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.

[0094] 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.

[0095] In S101, a first image including an object sequence is acquired.

[0096] In some embodiments, the object sequence may be a sequence formed by sequentially arranging any objects. A specific object type is not specially limited. For example, the first image is an image collected in a game place, and the object sequence may be tokens in a game in the game place. Alternatively, the first image 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.

[0097] The first image is at least one frame of image. The at least one frame of image is an image of which both size information and a pixel value satisfy certain conditions and which is obtained by size adjustment and pixel value normalization.

[0098] In some possible implementation modes, an acquired second image is preprocessed to obtain the first image that may be input to an object sequence recognition network. That is, S101 may be implemented through the following S111 and S112 (not shown in the figure).

[0099] In S111, a second image of which a picture includes the object sequence is acquired.

[00100] Here, the second image may be an image including appearance information of the object sequence. The second image may be an image collected by any collection device, or may be an image acquired from the Internet or another device or any frame in a video. For example, the second image is a frame of image which is acquired from a network and of which picture content includes the object sequence. Alternatively, the second image is a video segment of which picture content includes the object sequence, etc.

[00101] In S112, an image parameter of the second image is preprocessed based on a preset image parameter to obtain the first image.

[00102] In some possible implementation modes, the preset image parameter includes an image width, a height, an image pixel value, etc. First, size information of the second image is adjusted according to a preset size to obtain an adjusted image. The preset size is a preset width and a preset aspect ratio. For example, widths of multiple frames of second images are adjusted to the preset width according to the preset width in a unified manner. Then, pixel values of the adjusted image are normalized to obtain the first image. For example, for a second image of which a height is less than a preset height, an image region of which a height does not reach the preset height is filled with pixels, e.g., gray pixel values. As such, the size information is adjusted to make the aspect ratio in the size of the obtained first image the same, and deformations generated when the image is processed may be reduced.

[00103] In S102, the first image is input to an object sequence recognition network, and feature extraction is performed to obtain a feature sequence.

[00104] In some embodiments, supervision information in a training process of the

object sequence recognition network at least includes first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images. At least two frames of sample images in each sample image group include a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image. Image transformation may include one or more of scaling, rotation, brightness adjustment, etc. It is to be noted here that image transformation may not change a picture content of an image greatly and a transformed image has a picture content approximately the same as that of the image before transformed. Therefore, the robustness of the object sequence recognition network for changing a size, rotation angle, brightness and other dimensions of an image may be enhanced.

[00105] Alternatively, the object sequence recognition network is obtained by training based on a second loss for supervising a similarity between at least two frames of sample images with related or similar picture contents and a second loss set for supervising classes of sample objects in each frame of sample image. Multiple groups of sample images with related or similar picture contents are used for training in the training process of the object sequence recognition network.

[00106] The first input is input to the object sequence recognition network, and feature extraction is performed on the first image using a convolutional neural network part in the object sequence recognition network to obtain a feature map. 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 to facilitate subsequent classification of the object sequence in the first image. In some possible implementation modes, the feature map may be split according to any dimension of the feature map to obtain the feature sequence. In such case, a feature sequence of a frame of first image includes multiple features. 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.

[00107] In S103, a class of each object in the object sequence is determined based on

the feature sequence.

[00108] In some embodiments, a class of a feature in a feature sequence of each sample image in a group of sample images is predicted using a classifier in the object sequence recognition network, thereby obtaining a predicted probability of the feature sequence of each sample image. Class prediction may be performed on the feature in the feature sequence to obtain a class of each feature, the class of each feature being a class of an object corresponding to the feature. Therefore, feature sequences belonging to the same class are feature sequences corresponding to the same object, classes of features of the same class are a class of an object corresponding to the features of this class, and furthermore, the class of each object in the object sequence may be obtained. For example, if the total class number of the object sequence is n , the classes of the features in the feature sequence are predicted using a classifier with n class labels, thereby obtaining predicted probabilities that each feature in the feature sequence belongs to the n class labels. The predicted probabilities of the feature sequence are taken as a classification result of the feature sequence.

[00109] In some embodiments, the classification result of the feature sequence is processed according to a post-processing rule of a CTC function (for example, a final sequence recognition result is generated according to the probability of the sequence output by the CTC) to obtain the class of the object corresponding to each feature sequence, so that the class of each object in the object sequence in the first image may be predicted. A length of the sequence of the objects belonging to the same class may further be statistically obtained based on the class of the object corresponding to each feature sequence. The classification result of the feature sequence may represent a probability that the object corresponding to the feature sequence belongs to each classification label of the classifier. 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. As such, the class of each object may be obtained.

[00110] In the embodiment of the application, first, feature extraction is performed on the first image using the object sequence recognition network adopting supervision

information including the supervision on a similarity between a group of sample images and the supervision on classes of sample objects in the group of sample images to obtain the feature sequence. Then, class prediction is performed on each object in the object sequence, so that the obtained classification result of the object sequence is relatively accurate. Finally, the classification result of the object sequence is further processed to determine the classes of multiple objects. Accordingly, each object is classified using the network considering both the similarity between the images and the classes of the objects so that the consistency of the feature extraction and recognition results of similar images in the first image may be improved to improve the robustness, so that the object recognition accuracy is improved.

[00111] In some embodiments, the feature extraction of the first image 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 FIG. 2.

[00112] In S201, feature extraction is performed on the first image using a convolutional subnetwork in the object sequence recognition network to obtain a feature map.

[00113] In some embodiments, the object sequence recognition network is obtained by training based on a first loss for supervising a whole sample image and a second loss for supervising an object of each class in the sample image. Feature extraction is performed on the first image using a convolutional network part in the object sequence recognition network to obtain the feature map. The convolutional network part in the object sequence recognition network may be obtained by fine adjustment based on a network structure of a ResNet.

[00114] In some possible implementation modes, feature extraction is performed on the first image using a convolutional network obtained by stride adjustment in the object sequence recognition network, thereby obtaining a feature map of which a height is kept unchanged and a width is changed. That is, S201 may be implemented through the

following S211 to S213 (not shown in the figure).

[00115] In S211, the first image is down-sampled using the convolutional subnetwork in a length dimension of the first image in a first direction to obtain a first-dimensional feature.

[00116] In some possible implementation modes, a network structure of an adjusted ResNet is taken as a convolutional network for the feature extraction of the first image. The first direction is different from an arrangement direction of 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. For example, 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. In this manner, down-sampling in the length dimension of the first image in the first direction is implemented, and a length of the obtained feature map in the first direction is changed to a half of a length of the first image 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 first image is implemented, and a width of the obtained feature map is changed to a half of a width of the first image.

[00117] In S212, a feature in a length dimension of the first image in a second direction is extracted based on a length of the first image in the second direction to obtain a second-dimensional feature.

[00118] In some possible implementation modes, 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 first image in the second direction, namely the length of the first

image in the second direction is kept, and feature extraction is performed in the length direction of the first image in the second direction to obtain a second-dimensional feature the same as the length of the first image in the second direction.

[00119] In a specific example, the arrangement direction of the object sequence is the height direction. Height strides 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 a height dimension of the first image, namely the height of the first image is kept, and feature extraction is performed in the height direction of the first image to obtain a feature the same as the height of the first image.

[00120] In S213, the feature map is obtained based on the first-dimensional feature and the second-dimensional feature.

[00121] In some possible implementation modes, the first-dimensional feature is combined with the second-dimensional feature to form the feature map of the first image.

[00122] In S211 to S213, the first image is not down-sampled in the length dimension of the first image in the second direction to make a dimension of a dimensional feature in the second direction the same as that of the first image in the second direction, and the first image is down-sampled in the dimension in the first direction different from the arrangement direction of the object to change a length of a dimensional feature in the first direction to a half of the length of the first image in the first direction. As such, feature information of the first image in the dimension of the arrangement direction of the object sequence may be maximally retained. When the arrangement direction of the object sequence is the height direction, convolutional layers 3 and 4 of which last strides are (2, 2) in the ResNet are changed to convolutional layers of which strides are (1, 2), so that the first image is not down-sampled in the height dimension to make a dimension of a height-dimensional feature the same as the height of the first image, and the first image is down-sampled in the width dimension to change a width of a width-dimensional feature to a half of the width of the first image. As such, feature information of the first image in the height dimension may be maximally retained.

[00123] In S202, the feature map is split to obtain the feature sequence.

[00124] In some embodiments, the feature map is split based on dimension information of the feature map to obtain the feature sequence. The dimension information of the feature map includes a dimension in a first direction and a dimension in a second direction (e.g., a width dimension and a height dimension). The feature map is processed differently based on the two dimensions to obtain the feature sequence. For example, 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 using the object sequence recognition network obtained by training based on two loss functions, 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.

[00125] In some possible implementation modes, the feature map is pooled in the dimension in the first direction, and 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).

[00126] In S221, the feature map is pooled in the first direction to obtain a pooled feature map.

[00127] In some embodiments, average pooling is performed on the feature map in the dimension of the feature map in the first direction, and 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, the arrangement direction of the objects in the object sequence is the height direction, and the feature map is pooled in the width dimension in the dimension information to obtain the pooled feature map. A dimension of a first feature map is 2,048*40*16 (the channel dimension is 2,048, the height dimension is 40, and the width dimension is 16), and a 2,048*40*1 pooled feature map is obtained by average pooling in the width dimension.

[00128] In S222, the pooled feature map is split in the second direction to obtain the

feature sequence.

[00129] 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 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. In a specific example, the arrangement direction of the objects in the object sequence is the height direction, and the pooled feature map is split based on the height dimension to obtain the feature sequence. If the pooled feature map is $2,048 \times 40 \times 1$, the pooled 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 an image region in the height direction in the original first image. Accordingly, the feature map is split in the second direction the same as the arrangement direction of the objects after being pooled in the first direction different from the arrangement direction of the objects, so that the feature sequence may include more detail information of the first image in the second direction.

[00130] In some embodiments, the classification result of the feature sequence is further processed to predict the class of each object. That is, S103 may be implemented through the following S131 to S132 (not shown in the figure).

[00131] In S131, a class corresponding to each feature in the feature sequence is predicted using the classifier of the object sequence recognition network.

[00132] In some embodiments, the feature sequence is input to the classifier to predict the class corresponding to each feature in the feature sequence. For example, if the total class number of the object sequence is n , the class of the feature in the feature sequence is predicted using a classifier with n class labels, thereby obtaining a predicted probability that the feature in the feature sequence corresponds to each class label in the n class labels.

[00133] In S132, a class of each object in the object sequence is determined based on a prediction result of the class corresponding to each feature in the feature sequence.

[00134] In some embodiments, after the feature map is split, the feature sequence

includes multiple feature vectors of the image to be recognized in the dimension in the second direction, namely the feature vectors are part of features of the image to be recognized, may include all features of one or more object sequences or include part of features of an object sequence. As such, the classification result of the object corresponding to each feature in the feature sequence may be combined to accurately recognize the class of each object in the object sequence in the first image. The class of the object corresponding to each feature vector is predicted at first in the classification result of the feature sequence. Then, statistics may be made to the class of the object corresponding to each feature vector in the feature sequence to determine the class of each object in the first image. Accordingly, the classification result of the feature sequence is processed using a post-processing rule of a CTC loss function, so that the class of the object sequence in the image may be predicted more accurately.

[00135] In another embodiment, after the class of the object that each feature vector belongs to is determined, a length of the feature vectors of the objects belonging to the same class may be determined based on this. For example, the object sequence is a token sequence, and a token sequence length corresponding to features of tokens belonging to the same class is determined in the feature sequence. A class of a token is related to a face value of the token, a pattern of the token, a game that the token is suitable for, etc. A sequence length of features of objects of each class is indeterminate. Therefore, a fixed-length feature sequence is converted to a variable feature sequence length.

[00136] In some embodiments, the object sequence recognition network is configured to recognize the class of the object sequence. 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. 2B. FIG. 2B 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. 2B.

[00137] In S21, a sample image group is acquired.

[00138] In some embodiments, at least two frames of sample images in the sample image group include a frame of original sample image and at least one frame of image

obtained by performing image transformation on the original sample image. The original sample image includes a sample object sequence. The original sample image further includes class labeling information of sample objects. Similarities between picture contents of multiple frames of sample images in the sample image group are greater than a preset threshold. Alternatively, classes of sample objects labeled in multiple frames of sample images in the sample image group are the same.

[00139] In S22, the sample images in the sample image group are input to an object sequence recognition network to be trained, and feature extraction is performed to obtain a sample feature sequence of each sample image in the sample image group.

[00140] In some embodiments, the sample images in the sample image group are preprocessed at first to make sizes of the sample images in the sample image group the same. Then, feature extraction is performed on a processed sample image group to obtain a sample feature sequence of each sample image.

[00141] In some possible implementation modes, a paired image of a collected original sample image is created at first through the sample image (i.e., a first sample image). Then, a sample image pair is formed, data enhancement is performed on the paired image, and the original sample image and an enhanced image are combined as a sample image group. That is, S21 may be implemented through the following S231 to S234 (not shown in the figure).

[00142] In S231, a first sample image where a class of a sample object in a picture is labeled is acquired.

[00143] Here, image collection may be performed on a scene with the sample object using an image collection device to obtain the first sample image, and the class of the sample object in the first sample image is labeled. The first sample image may be one or more frames of images.

[00144] In S232, at least one second sample image is determined based on a picture content of the first sample image.

[00145] In some possible implementation modes, multiple images with relatively high

similarities with the picture content are generated at first according to the picture content of the first sample image, namely multiple second sample images are obtained.

[00146] In S233, data enhancement is performed on the at least one second sample image to obtain at least one third sample image.

[00147] In some embodiments, data enhancement is performed on the generated second sample image. For example, horizontal flipping, random pixel disturbance addition, image resolution or brightness adjustment, and the like are performed on the second sample image to obtain the third sample image.

[00148] In a specific example, the picture content of the first sample image is copied to obtain the second sample image. Then, data enhancement is performed on the second sample image, for example, a resolution, brightness or the like of the second sample image is adjusted, to obtain the third sample image.

[00149] In S234, the sample image group is obtained based on the first sample image and the at least one third sample image.

[00150] In some embodiments, size information adjustment, pixel value normalization, and the like are performed on the first sample image and multiple third sample images to implement the unification of sample data. At least two frames of images unified in such a manner are taken as a sample image group. As such, paired images with similar picture contents are created through each frame of first sample image to make it easy to subsequently improve the feature extraction consistency of similar images.

[00151] In some possible implementation modes, sample images in a group of sample images are preprocessed, data enhancement is performed on preprocessed images, and the preprocessed images and enhanced images are combined as a final group of sample images. That is, S234 may be implemented through the following process.

[00152] First, image parameters of the at least one third sample image and the first sample image are preprocessed based on a preset image parameter to obtain at least two frames of third sample images.

[00153] In some possible implementation modes, a process of preprocessing the image parameters of the third sample image and the first sample image is similar to that of preprocessing the second image. That is, size information of the third sample image and the first sample image is adjusted at first according to a preset size to obtain adjusted images. Then, pixel values of the adjusted images are normalized to obtain at least two frames of third sample images. As such, widths of multiple frames of sample original images are adjusted to the preset width according to the preset width in a unified manner. For a sample original image of which a height is less than a preset height, an image region of which a height does not reach the preset height is filled with pixels, e.g., gray pixel values. As such, the size information is adjusted to make aspect ratios in sizes of the multiple frames of sample images in the obtained sample image group the same, and errors generated in the training process of the network may be reduced.

[00154] Then, data enhancement is performed on the at least two frames of third sample images to obtain the sample image group.

[00155] Here, data enhancement includes random flipping, random clipping, random aspect ratio fine adjustment, random rotation, and other operations. Therefore, random flipping, random clipping, random aspect ratio fine adjustment, random rotation and other operations may be performed on the multiple frames of adjusted images to obtain a richer group of sample images. Accordingly, the multiple frames of images of which the sizes are unified are combined as the sample image group. Therefore, sample images may be enriched, and the overall robustness of the network to be trained may be improved.

[00156] In S23, a class of a sample object sequence in each sample image is predicted based on the sample feature sequence of each sample image.

[00157] In some embodiments, a class of each sample feature in the sample feature sequence of each sample image in the sample image group is predicted at first to obtain a classification result of each sample feature. Then, a sample object classification result of each sample image is determined based on the classification result of each sample feature. As such, the sample feature sequence of each sample image in the sample image group is input to a classifier of the object sequence recognition network to be trained, and class prediction is performed to obtain the sample classification result of the sample feature

sequence of each sample image. The sample object classification result of each sample image includes the classification result of each sample feature in the sample feature sequence of the sample image.

[00158] In some possible implementation modes, all classes of the sample objects are analyzed, and classification labels of the classifier are set, so that the sample classification result of each sample feature sequence is predicted. That is, S23 may be implemented through the following process.

[00159] First, total classes of the sample objects in the sample image set are determined.

[00160] Here, all classes of the sample objects in a scene of the sample images are analyzed. For example, in a game scene, the sample object is token, and all possible classes of tokens are determined as total classes of the tokens.

[00161] Then, classification labels of the classifier of the object sequence recognition network to be trained are determined based on the total classes of the sample objects.

[00162] Here, the classification labels of the classifier are set according to the total classes of the sample objects, and then the classifier may predict probabilities that the sample objects in the sample image belong to any class.

[00163] Finally, class prediction is performed on the sample objects in the sample feature sequence of each sample image in the sample image group using the classifier with the classification labels to obtain the sample object classification result of each sample image.

[00164] Here, a class of an object in the sample feature sequence of each sample image may be predicted using a classifier with multiple classification labels to obtain the sample classification result of the sample feature sequence of the sample image. The class that the object in the sample feature sequence most probably belongs to may be determined based on the sample classification result. Accordingly, the total classes of the objects are analyzed, and the classification labels that the classifier has are set, so that the classes of the objects in the sample feature sequence may be predicted more accurately.

[00165] In S24, a first loss for supervising the class of the sample object sequence in each sample image is determined based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set.

[00166] In some embodiments, a CTC loss is adopted as the first loss. For each sample image in the sample image group, the first 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 class labeling information of the sample object in the sample image as an input of the CTC loss to predict the class of the sample object corresponding to each feature in the sample feature sequence of the sample image. As such, the first loss set may be obtained based on the group of sample images.

[00167] In S25, a second loss for supervising a similarity between the at least two frames of sample images is determined based on the sample feature sequences of the at least two frames of sample images in the sample image group.

[00168] In some embodiments, a pair loss is adopted as the second loss. For example, an implementation mode of the pair loss may be selecting from losses for measuring distribution differences, e.g., a Layer 2 (L2) loss, a cos loss, and a Kullback-Leibler divergence loss. Similarities between different sample images in a group of sample image are predicted taking a similarity truth value between a sample feature sequence of each sample image and the sample image as an input of the second loss.

[00169] In S26, a network parameter of the object sequence recognition network to be trained is adjusted using the first loss set and the second loss such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.

[00170] Here, a class representing each sample object in the classification result and truth value information of each sample object may be compared to determine the first loss set. Similarities between different sample images and similarity truth values between different sample images may be compared to determine the second loss. The first losses and the second loss 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.

[00171] Through S21 to S26, the first loss set for supervising whole sequences and the second loss for supervising the similarity between a similarity between images in a group of sample images are introduced to the object sequence recognition network to be trained based on the sample image group formed by paired images, so that the feature extraction consistency of similar images may be improved, and an overall class prediction effect of the network may be improved.

[00172] In some embodiments, the feature extraction of the sample image is implemented using a convolutional subnetwork in the object sequence recognition network to be trained, thereby obtaining the sample feature sequence. That is, S22 may be implemented through the following S241 and S242 (not shown in the figure).

[00173] In S241, feature extraction is performed on each sample image using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image.

[00174] In some embodiments, feature extraction is performed on each sample image in the sample image group using a convolutional network obtained by finely adjusting the structure of a ResNet as the convolutional subnetwork of the recognition network to be trained to obtain the sample feature map.

[00175] In some possible implementation modes, feature extraction is performed on the sample image using a convolutional subnetwork obtained by stride adjustment in the object sequence recognition network, thereby obtaining a sample feature map which of which information in a dimension in a first direction is kept unchanged and information in a dimension in a second direction is changed. That is, S241 may be implemented through the following operations.

[00176] In a first step, each sample image 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.

[00177] Here, the first direction is different from an arrangement direction of sample

objects in the sample object sequence. An implementation process of the first step is similar to that of S211. When the arrangement direction of the sample object sequence is a stacking height direction, the sample image is down-sampled in a width dimension of the sample image to obtain a first-dimensional sample feature. It is set that width strides in last strides of convolutional layers 3 and 4 of the convolutional subnetwork are kept at 2 and unchanged, and height strides are changed from 2 to 1.

[00178] In a second step, feature extraction is performed in a length dimension of each sample image in a second direction based on a length of each sample image in the second direction to obtain a second-dimensional sample feature.

[00179] Here, an implementation process of the second step is similar to that of S212. When the arrangement direction of the sample object sequence is a stacking height direction, feature extraction is performed based on a height of the sample image in a height dimension of the sample image to obtain a second-dimensional sample feature. For example, it is set that heights in the last strides of convolutional layers 3 and 4 of the convolutional subnetwork are changed from 2 to 1. In such case, down-sampling is not performed in the height dimension of the sample image, namely the second-dimensional sample feature of the sample image is kept.

[00180] In a third step, the sample feature map of each sample image is obtained based on the first-dimensional sample feature and the second-dimensional sample feature.

[00181] Here, the first-dimensional sample feature and the second-dimensional sample feature are combined to form the sample feature map of the sample image.

[00182] Through the first step to the third step, when the arrangement direction of the sample object sequence is the stacking height direction, convolutional layers 3 and 4 of which last strides are (2, 2) in the ResNet are changed to convolutional layers of which strides are (1, 2) to form the convolutional subnetwork of the object sequence recognition network to be trained. Therefore, feature information of each sample image in the dimension in the arrangement direction may be maximally retained.

[00183] In S242, the sample feature map of each sample image is split to obtain the sample feature sequence of each sample image.

[00184] Here, an implementation process of S242 is similar to that of S202. That is, each sample feature map is processed differently based on the dimension in the first direction and the dimension in the second direction to obtain the sample feature sequence. For example, the sample feature map is pooled in the dimension in the first direction, and is split into multiple feature vectors in the dimension in the second direction to form the sample feature sequence. As such, the obtained sample feature sequence may retain more dimensional features of the sample object in the arrangement direction, and the training accuracy of the network may be improved.

[00185] In some possible implementation modes, the sample feature map is pooled in the dimension in the second direction, and is split in the dimension in the first direction to obtain the sample feature sequence. That is, S242 may be implemented through the following operations.

[00186] In a first step, the sample feature map is pooled in the first direction to obtain a pooled sample feature map.

[00187] Here, an implementation process of the first step is similar to that of S221. That is, average pooling is performed on the sample feature map in the dimension of the sample feature map in the first direction, and the dimension of the sample feature map in the second direction and a channel dimension are kept unchanged, to obtain the pooled sample feature map.

[00188] In a second step, the pooled sample feature map is split in the second direction to obtain the sample feature sequence.

[00189] Here, an implementation process of the second step is similar to that of S222. That is, the pooled sample feature map is split in the dimension of the sample feature map in the second direction to obtain the sample feature sequence. For example, if the dimension of the sample feature map in the second direction is 40, the pooled sample feature map is split into 40 vectors to form a sample feature sequence. Accordingly, the sample feature map is split in the dimension in the second direction after being pooled in the dimension in the second direction, so that the sample feature sequence may retain more detailed information of the sample image in the dimension in the second direction.

- [00190] In some embodiments, dynamic weighted fusion is performed on the first loss set and the second loss to improve the object sequence recognition performance of the object sequence recognition network to be trained. That is, S26 may be implemented through the following S261 and S262.
- [00191] In S261, weighted fusion is performed on the first loss set and the second loss to obtain a total loss.
- [00192] In some embodiments, the first loss set and the second loss are weighted using different dynamic weights, and a first loss set and second loss which are obtained by weighted adjustment are fused to obtain the total loss.
- [00193] In some embodiments, adjustment parameters are set for the first losses and the second loss to improve the object sequence recognition performance of the object sequence recognition network to be trained. That is, S261 may be implemented through the following operations.
- [00194] In a first step, a class supervision weight corresponding to the sample image group is determined based on the number of the sample images in the sample image group.
- [00195] In some embodiments, the class supervision weight corresponding to the sample image group is determined for the group of sample images. That is, weights for first losses corresponding to each sample image in the same group may be the same. The number of class supervision weights is the same as that of the sample images. 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 $1/n$.
- [00196] In a second step, the first losses in the first loss set of the sample image group are fused based on the class supervision weight and a first preset weight to obtain a third loss.
- [00197] In some embodiments, the class supervision weight and the first preset weight

may both be assigned to the first losses in the first loss set, and multiple first losses assigned with the parameters are summed to obtain the third loss.

[00198] In some possible implementation modes, the multiple first losses are fused at first using the class supervision weight, and then the fused loss is adjusted using the first preset weight, thereby obtaining the third loss. That is, the second step may be implemented through the following operations.

[00199] In step A, the class supervision weight is assigned to each first loss in the first loss set to obtain an updated loss set including at least two updated losses.

[00200] In some possible implementation modes, if the number of the sample images in the sample image group is 4, the class supervision weight is 0.25, and the class supervision weight 0.25 is multiplied by each first loss to obtain an updated loss.

[00201] In step B, the updated losses in the updated loss set are fused to obtain a fused loss.

[00202] In some possible implementation modes, the updated losses in the updated loss set are added to obtain the fused loss.

[00203] In step C, the fused loss is adjusted using the first preset weight to obtain the third loss.

[00204] In some possible implementation modes, a ratio of the first preset weight to a second preset weight is set to be 1:10. The first preset weight is multiplied by the fused loss to obtain the third loss. Accordingly, 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.

[00205] In a third step, the second loss is adjusted using a second preset weight to obtain a fourth loss.

[00206] In some embodiments, the second preset weight is assigned to the second loss to implement the adjustment of the second loss. For example, the second preset weight is multiplied by the second loss to obtain the fourth loss. The first preset weight is less than

the second preset weight. The first preset weight and the second preset weight satisfy a certain proportional relationship.

- [00207] In S254, the total loss is determined based on the third loss and the fourth loss.
- [00208] In some embodiments, the third loss and the fourth loss are added to obtain the total loss of the object sequence recognition network to be trained.
- [00209] In S262, the network parameter of the object sequence recognition network to be trained is adjusted based on the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.
- [00210] In the embodiment of the application, 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 the feature extraction consistency of similar images may be improved, and the prediction effect of the whole network may be improved.
- [00211] An exemplary application of the embodiment of the application to a practical application scene will be described below. Taking a game place as an example of the application scene, descriptions will be made with the recognition of an object (e.g., a token) in the game place as an example.
- [00212] 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.
- [00213] However, for the recognition of a token sequence in the game place, the 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 chips based on a deep learning method is not so good.

[00214] 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 chip recognition, so that the feature extraction consistency of similar images may be improved, and the chip recognition accuracy may further be improved.

[00215] FIG. 3 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. 3. A framework of the object sequence recognition network includes a paired data construction module, a feature extraction module 303, a classifier, and a loss module.

[00216] The paired data construction module is configured to construct a corresponding paired image 302 for each frame of image 301 in training images to obtain a sample image set.

[00217] In some possible implementation modes, a data enhancement process is performed on the training images. For example, horizontal flipping, random pixel disturbance addition, image resolution adjustment, image brightness adjustment and other operations are performed on the training images. The corresponding paired image is constructed for each frame of image without changing the labeling of a token sequence in the image, and a subsequent process is performed in pairs.

[00218] In a specific example, a frame of image is copied at first, and then data enhancement is performed on a copied image to obtain a paired image of the image.

[00219] After the corresponding paired image is constructed for each frame of image, a frame of sample image is preprocessed, including adjusting a size of the image with an aspect ratio kept unchanged, normalizing pixel values of the image, and other operations. The operation of adjusting the size of the image with the aspect ratio kept unchanged refers to adjusting widths of multiple frames of sample images to be the same. In order to reduce great deformations of the multiple frames of images generated if aspect ratios of the multiple frames of images are not adjusted to be the same because tokens in the input images are different in number and the aspect ratios of the images are greatly different, in

the embodiment of the application, 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 the sample image set, a data enhancement operation is performed on processed sample images. For example, random flipping, random clipping, random aspect ratio fine adjustment, random rotation and other operations are performed on the processed sample images. As such, the overall robustness of the network to be trained may be improved.

[00220] The feature extraction module 303 performs feature extraction on processed sample images to obtain feature sequences 303. Feature extraction is performed to obtain a sample feature sequence 31 after the image 301 is processed. Feature extraction is performed to obtain a sample feature sequence 32 after the paired image 302 is processed.

[00221] In some possible implementation modes, high-layer features of the input sample images 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 are (2, 2) of convolutional layers 3 and 4 in the network structure of the ResNet are changed to strides (1, 2). As such, a feature map is not down-sampled in a height dimension, and is down-sampled in a width dimension to halve an original width. Therefore, feature information in the height dimension may be maximally retained. Then, a splitting operation is performed on the feature map, namely the feature map extracted by the convolutional neural network is split into a plurality of feature sequences to facilitate subsequent calculation of the classifier and a loss function. When the feature map is split, average pooling is performed in a width direction of the feature map, and not 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.

[00222] In a specific example, if the sample image includes multiple tokens, as shown

in FIG. 4, the feature sequence is obtained by division according to a height dimension of the image 401. A feature sequence includes a feature of less than or equal to one token.

[00223] The classifier adopts an n-classifier, and performs token class prediction on the feature sequence to obtain a predicted probability of each feature sequence.

[00224] Here, n is the total number of token classes.

[00225] For the feature sequence obtained by the convolutional network, the loss module determines a feature similarity of the paired images using the pair loss 304 and supervises the network for an optimization purpose of improving the similarity. For predicted probabilities of all feature sequence classes, a prediction result of a feature sequence of each image in a pair of images is supervised using a CTC loss 305 and a CTC loss 306 respectively.

[00226] In some possible implementation modes, the pair loss 304, the CTC loss 305 and the CTC loss 306 form a total loss 307: $L = \alpha(0.5L_{ctc1} + 0.5L_{ctc2}) + \beta L_{pair}$, where values of α and β may be set as $\alpha : \beta = 1 : 10$.

[00227] 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.

[00228] 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 class recognition accuracy may be improved to finally improve the overall recognition result, particularly in a scene with a long token sequence.

[00229] An embodiment of the application provides an object sequence recognition apparatus. FIG. 5A is a structure composition diagram of an object sequence recognition

apparatus according to an embodiment of the application. As shown in FIG. 5A, the object sequence recognition apparatus 500 includes a first acquisition module 501, a first extraction module 502, and a first determination module 503.

[00230] The first acquisition module 501 is configured to acquire a first image including an object sequence.

[00231] The first extraction module 502 is configured to input the first image to an object sequence recognition network and perform feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least including first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image.

[00232] The first determination module 503 is configured to determine a class of each object in the object sequence based on the feature sequence.

[00233] In some embodiments, the first extraction module 502 includes a first extraction submodule and a first splitting submodule.

[00234] The first extraction submodule is configured to perform feature extraction on the first image using a convolutional subnetwork in the object sequence recognition network to obtain a feature map.

[00235] The first splitting submodule is configured to split the feature map to obtain the feature sequence.

[00236] In some embodiments, the first extraction submodule includes a first down-sampling unit, a first extraction unit, and a first determination unit.

[00237] The first down-sampling unit is configured to down-sample the first image using the convolutional subnetwork in a length dimension of the first image 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.

[00238] The first extraction unit is configured to extract a feature in a length dimension of the first image in a second direction based on a length of the first image in the second direction to obtain a second-dimensional feature.

[00239] The first determination unit is configured to obtain the feature map based on the first-dimensional feature and the second-dimensional feature.

[00240] In some embodiments, the first splitting submodule includes a first pooling unit and a first splitting unit.

[00241] The first pooling unit is configured to pool the feature map in the first direction to obtain a pooled feature map.

[00242] The first splitting unit is configured to split the pooled feature map in the second direction to obtain the feature sequence.

[00243] In some embodiments, the first determination module 503 includes a first prediction submodule and a first determination submodule.

[00244] The first prediction submodule is configured to predict a class corresponding to each feature in the feature sequence using a classifier of the object sequence recognition network.

[00245] The first determination submodule is configured to determine the class of each object in the object sequence based on a prediction result of the class corresponding to each feature in the feature sequence.

[00246] An embodiment of the application provides an apparatus for training an object sequence recognition network. FIG. 5B 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. 5B, the apparatus 510 for training an object sequence recognition network includes a second acquisition module 511, a second extraction module 512, a first prediction module 513, a second determination module 514,

a third determination module 505, and a first adjustment module 516.

[00247] The second acquisition module 511 is configured to acquire a sample image group, at least two frames of sample images in the sample image group including a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image including a sample object sequence.

[00248] The second extraction module 512 is configured to input the sample images in the sample image group to an object sequence recognition network to be trained and perform feature extraction to obtain a sample feature sequence of each sample image in the sample image group.

[00249] The first prediction module 513 is configured to predict a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image.

[00250] The second determination module 514 is configured to determine a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set.

[00251] The third determination module 515 is configured to determine a second loss for supervising a similarity between the at least two frames of sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group.

[00252] The first adjustment module 516 is configured to adjust a network parameter of the object sequence recognition network to be trained using the first loss set and the second loss such that a loss of a classification result output by an adjusted object sequence recognition network satisfies a convergence condition.

[00253] In some embodiments, the second acquisition module 511 includes a first acquisition submodule, a second determination submodule, a first enhancement submodule, and a third determination submodule.

- [00254] The first acquisition submodule is configured to acquire a first sample image where a class of a sample object in a picture is labeled.
- [00255] The second determination submodule is configured to determine at least one second sample image based on a picture content of the first sample image.
- [00256] The first enhancement submodule is configured to perform data enhancement on the at least one second sample image to obtain at least one third sample image.
- [00257] The third determination submodule is configured to obtain the sample image group based on the first sample image and the at least one third sample image.
- [00258] In some embodiments, the second extraction module 512 includes a second extraction submodule and a second splitting submodule.
- [00259] The second extraction submodule is configured to perform feature extraction on each sample image using a convolutional subnetwork in the object sequence recognition network to be trained to obtain a sample feature map of each sample image.
- [00260] The second splitting submodule is configured to split the sample feature map of each sample image to obtain the sample feature sequence of each sample image.
- [00261] In some embodiments, the second extraction submodule includes a second down-sampling unit, a second extraction unit, and a second determination unit.
- [00262] The second down-sampling unit is configured to down-sample each sample image 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.
- [00263] The second extraction unit is configured to perform feature extraction in a length dimension of each sample image in a second direction based on a length of each sample image in the second direction to obtain a second-dimensional sample feature.
- [00264] The second determination unit is configured to obtain the sample feature map of each sample image based on the first-dimensional sample feature and the second-

dimensional sample feature.

[00265] In some embodiments, the second splitting submodule includes a second pooling unit and a second splitting unit.

[00266] The second pooling unit is configured to pool the sample feature map in the first direction to obtain a pooled sample feature map.

[00267] The second splitting unit is configured to split the pooled sample feature map in the second direction to obtain the sample feature sequence.

[00268] In some embodiments, the first adjustment module 516 includes a first fusion submodule and a first adjustment submodule.

[00269] The first fusion submodule is configured to perform weighted fusion on the first loss set and the second loss to obtain a total loss.

[00270] The first adjustment submodule is configured to adjust the network parameter of the object sequence recognition network to be trained based on the total loss such that the loss of the classification result output by the adjusted object sequence recognition network satisfies the convergence condition.

[00271] In some embodiments, the first fusion submodule includes a third determination unit, a first fusion unit, a first adjustment unit, and a fourth determination unit.

[00272] The third determination unit is configured to determine a class supervision weight corresponding to the sample image group based on the number of the sample images in the sample image group.

[00273] The first fusion unit is configured to fuse the first losses in the first loss set of the sample image group based on the class supervision weight and a first preset weight to obtain a third loss.

[00274] The first adjustment unit is configured to adjust the second loss using a second preset weight to obtain a fourth loss.

- [00275] The fourth determination unit is configured to determine the total loss based on the third loss and the fourth loss.
- [00276] In some embodiments, the first fusion unit includes a first assignment subunit, a first fusion subunit, and a first adjustment subunit.
- [00277] The first assignment subunit is configured to assign the class supervision weight to each first loss in the first loss set to obtain an updated loss set including at least two updated losses.
- [00278] The first fusion subunit is configured to fuse the updated losses in the updated loss set to obtain a fused loss.
- [00279] The first adjustment subunit is configured to adjust the fused loss using the first preset weight to obtain the third loss.
- [00280] It is to be noted that the 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.
- [00281] It is to be noted that, in the embodiments of the application, the object sequence recognition method may also be stored in a computer-readable storage medium when being implemented in form of 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 terminal, a server, 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.

[00282] An embodiment of the application also provides a computer program product including a computer-executable instruction which may be executed to implement the object sequence recognition method and the method for training an object sequence recognition network in the embodiments of the application.

[00283] An embodiment of the application also provides a computer storage medium having stored therein a computer-executable instruction which is executed by a processor to implement the object sequence recognition method and the method for training an object sequence recognition network in the abovementioned embodiments.

[00284] An embodiment of the application provides a computer device. FIG. 6 is a composition structure diagram of a computer device according to an embodiment of the application. As shown in FIG. 6, the computer device 600 includes a processor 601, at least one communication bus, a communication interface 602, at least one external communication interface, and a memory 603. The communication interface 602 is configured to implement connections and communications between these components. The communication interface 602 may include a display screen. The external communication interface may include a standard wired interface and wireless interface. The processor 601 is configured to execute an image processing 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.

[00285] The above descriptions about the embodiments of the object sequence 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 sequence 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.

[00286] 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.

[00287] 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.

[00288] 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.

[00289] 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.

[00290] 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 a first image comprising an object sequence;

inputting the first image to an object sequence recognition network, and performing feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least comprising first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image; and

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

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

performing feature extraction on the first image 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 first image using a convolutional subnetwork in the object sequence recognition network to obtain a feature map comprises:

down-sampling the first image using the convolutional subnetwork in a length

dimension of the first image 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 first image in a second direction based on a length of the first image 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. The method of any one of claims 1-4, wherein the determining a class of each object in the object sequence based on the feature sequence comprises:

predicting a class corresponding to each feature in the feature sequence using a classifier of the object sequence recognition network; and

determining the class of each object in the object sequence based on a prediction result of the class corresponding to each feature in the feature sequence.

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

acquiring a sample image group, at least two frames of sample images in the sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original

sample image comprising a sample object sequence;

inputting the sample images in the sample image group to an object sequence recognition network to be trained, and performing feature extraction to obtain a sample feature sequence of each sample image in the sample image group;

predicting a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image;

determining a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set;

determining a second loss for supervising a similarity between the at least two frames of sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group; and

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

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

acquiring a first sample image where a class of a sample object in a picture is labeled;

determining at least one second sample image based on a picture content of the first sample image;

performing data enhancement on the at least one second sample image to obtain at least one third sample image; and

obtaining the sample image group based on the first sample image and the at least one third sample image.

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

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

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

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

down-sampling each sample image 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;

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

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

10. The method of claim 9, wherein the splitting the sample feature map of each sample image to obtain the sample feature sequence of each sample image comprises:

pooling the sample feature map in the first direction to obtain a pooled sample feature

map; and

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

11. The method of any one of claims 6-10, wherein the adjusting a network parameter of the object sequence recognition network to be trained using the first loss set and the second loss 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 set and the second loss to obtain a total loss; and

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

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

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

fusing first losses in the first loss set of the sample image group based on the class supervision weight and a first preset weight to obtain a third loss;

adjusting the second loss using a second preset weight to obtain a fourth loss; and

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

13. The method of claim 12, wherein the fusing first losses in the first loss set of the sample

image group based on the class supervision weight and a first preset weight to obtain a third loss comprises:

assigning the class supervision weight to each first loss in the first 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 first preset weight to obtain the third loss.

14. A computer storage medium, having a computer-executable instruction stored thereon, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire a first image comprising an object sequence;

input the first image to an object sequence recognition network, and perform feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least comprising first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image; and

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

15. A computer storage medium, having a computer-executable instruction stored thereon, wherein when executed by a processor, the computer-executable instruction is configured to:

acquire a sample image group, at least two frames of sample images in the sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image comprising a sample object sequence;

input the sample images in the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain a sample feature sequence of each sample image in the sample image group;

predict a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image;

determine a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set;

determine a second loss for supervising a similarity between the at least two frames of sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group; and

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

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

acquire a first image comprising an object sequence;

input the first image to an object sequence recognition network, and perform feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least comprising first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image; and

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

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

acquire a sample image group, at least two frames of sample images in the sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image comprising a sample object sequence;

input the sample images in the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain a sample feature sequence of each sample image in the sample image group;

predict a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image;

determine a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set;

determine a second loss for supervising a similarity between the at least two frames of sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group; and

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

18. 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 first image comprising an object sequence;

input the first image to an object sequence recognition network, and perform feature extraction to obtain a feature sequence, supervision information in a training process of the object sequence recognition network at least comprising first supervision information of a class of a sample object sequence in each sample image of a sample image group and second supervision information of a similarity between at least two frames of sample images, and at least two frames of sample images in each sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image; and

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

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 a sample image group, at least two frames of sample images in the sample image group comprising a frame of original sample image and at least one frame of image obtained by performing image transformation on the original sample image, and the original sample image comprising a sample object sequence;

input the sample images in the sample image group to an object sequence recognition network to be trained, and perform feature extraction to obtain a sample feature sequence of each sample image in the sample image group;

predict a class of a sample object sequence in each sample image based on the sample feature sequence of each sample image;

determine a first loss for supervising the class of the sample object sequence in each sample image based on the sample feature sequence of each sample image in the sample image group to obtain a first loss set;

determine a second loss for supervising a similarity between the at least two frames of

sample images based on the sample feature sequences of the at least two frames of sample images in the sample image group; and

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