APPARATUS FOR RECOGNIZING OBJECTS BASED ON ARTIFICIAL INTELLIGENCE (AI) AND METHOD THEREFOR

TECHNICAL FIELD

The present disclosure relates to an object recognition technology, and more particularly, to an apparatus for recognizing objects based on artificial intelligence (AI) and a method therefor.

BACKGROUND

Data re-identification is a process of matching anonymized data ("de-identified data") with previously published information or ancillary data to identify individuals who belong to data. This raises concerns that institutions that handle personal information, healthcare providers, and financial institutions may disclose collected data after data has undergone a de-identification process. The de-identification process involves masking, generalization, and deletion of direct and indirect identifiers. However, there is no universal definition for this process. Even if information in the public domain appears to be anonymized, it can be re-identified using a combination of other available data and basic computer science techniques. However, there is no universal definition for this process. Even if information in the public domain appears to be anonymized, the information can be re-identified using a combination of other available data and basic computer science techniques.

SUMMARY

In view of the above, the present disclosure provides an apparatus for recognizing objects based on artificial intelligence (AI) and a method therefor.

According to embodiments of the present disclosure, a method of recognizing objects

includes: extracting, by a source detection unit, a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality of consecutive frames of each of a plurality of image sources; deriving, by a source feature detection unit, a plurality of source feature vectors representing features of the plurality of representative images through a feature extraction model which is a deep learning model; deriving, by a target feature detection unit, a target feature vector representing a feature of the query image through the feature extraction model when a query image, which is the image of the human object to be re-identified, is input; deriving, by a re-identification unit, a similarity between the human object to be re-identified and the human object of the plurality of representative images by calculating a similarity between the target feature vector and the plurality of source feature vectors by a re-identification unit; and providing, by the re-identification unit, representative images in order of the derived high similarity.

The deriving of the target feature vector may include: deriving, by a backbone network of the feature extraction model, a backbone feature vector through a weight calculation for the query image; generating, by a first pooling layer of the feature extraction model, a first pooling feature vector through a pooling operation on the backbone feature vector; generating, by a second pooling layer of the feature extraction model, a second pooling feature vector through the pooling operation on the backbone feature vector; generating, by a first batch normalization layer of the feature extraction model, a first batch normalization feature vector through batch normalization using a moving average and variance of the first pooling feature vector; generating, by a second batch normalization layer of the feature extraction model, a second batch normalization feature vector through batch normalization using a moving average and variance of the second pooling feature vector; and performing, by a fully-connected layer of the feature extraction model, a weight calculation using an activation function on the first batch normalization feature vector and the second

batch normalization feature vector to derive the target feature vector.

The detecting of the plurality of representative images may include: detecting, by a detection module of the source detection unit, the plurality of human objects from the plurality of consecutive frames for each image source through a bounding box representing a region occupied by the human object using an object detection model, which is the deep learning model, to assign a source identifier; assigning, by a tracking module of the source detection unit, the same tracking identifier to the bounding box determined to represent the same human object over the plurality of consecutive frames to track the human object; and extracting, by a selection module of the source detection unit, an image within any one bounding box from a plurality of bounding boxes to which the same tracking identifier is assigned to generate the representative image.

In the tracking of the human object, the tracking module may compare areas and positions of a bounding box of a current frame and a bounding box of a previous frame among the plurality of consecutive frames to determine, as the same human object as the bounding box of the previous frame, the bounding box of the current frame in which an area loss representing an area in which the bounding box of the previous frame and the bounding box of the current frame do not overlap is less than a preset first threshold value and a position loss representing a difference between a position of center coordinates of the bounding box of the previous frame and a position of center coordinates of the bounding box of the current frame is less than a preset second threshold value, and assign the same identifier.

In the tracking of the human object, the tracking module may derive the area loss depending on the following Equation.

[Equation]

$$La = 1 - \frac{2\sum_{i}(p_i + q_i) + \varepsilon}{\sum_{i}(p_i^2 + q_i^2) + \varepsilon}$$

Here, La may denote the area loss, i may denote an index of a pixel, q may denote the bounding box of the current frame, p may denote the bounding box of the previous frame, p and q may have a value of 1 when present in the pixel indicated by the index of the pixel, and have a value of 0 when not present, and ε may denote an adjustment constant.

In the tracking of the human object, the tracking module may derive the position loss depending on the following Equation.

[Equation]

$$Ld = \| R - S \|_2^2$$

Here, Ld may denote the position loss, R may denote a position vector of the center coordinates of the bounding box of the previous frame, and S may denote a position vector of the center coordinates of the bounding box of the current frame.

The deriving of the similarity may include: performing, by the re-identification unit, clustering on the target feature vector and the plurality of source feature vectors to derive a plurality of clusters; deriving, by the re-identification unit, a convex hull from a cluster including the target feature vector among the plurality of clusters; setting, by the re-identification unit, a center of gravity of a polygon constituting the convex hull as a corresponding cluster center; setting, by the re-identification unit, as a search area, an inner side of a circle having a radius of a distance from a cluster center to a target feature vector based on the cluster center; and calculating, by the re-identification unit, a similarity

according to a distance between the target feature vector and the source feature vector within the search area.

The deriving of the plurality of clusters may include: selecting, by the re-identification unit, center vectors of a predetermined number of clusters including the target feature vector and target feature vectors among the plurality of source feature vectors; and performing, by the re-identification unit, clustering on the target feature vector and the plurality of source feature vectors while reselecting the center vector until all the source feature vectors are assigned to a cluster with a smallest distance from the center vector of any one cluster to derive the plurality of clusters.

An apparatus for recognizing objects may include: a source detection unit configured to extract a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality of consecutive frames of each of a plurality of image sources; a source feature detection unit configured to derive a plurality of source feature vectors representing features of the plurality of representative images through a feature extraction model which is a deep learning model; a target feature detection unit configured to derive a target feature vector representing a feature of the query image through the feature extraction model when a query image, which is the image of the human object to be re-identified, is input; and a re-identification unit configured to calculate a similarity between the target feature vector and the plurality of source feature vectors to derive a similarity between the human object to be re-identified and the human object of the plurality of representative images, and provide representative images in order of the derived high similarities.

The feature extraction model may include: a backbone network configured to derive a backbone feature vector through a weight calculation for the query image; a first pooling layer configured to generate a first pooling feature vector through a pooling operation for the

backbone feature vector; a second pooling layer configured to generate a second pooling feature vector through the pooling operation for the backbone feature vector; a first batch normalization layer configured to generate a first batch normalization feature vector through batch normalization using a moving average and variance of the first pooling feature vector; a second batch normalization layer configured to generate a second batch normalization feature vector through batch normalization using a moving average and variance of the second pooling feature vector; and a fully-connected layer configured to perform a weight calculation using an activation function on the first batch normalization feature vector and the second batch normalization feature vector to derive the target feature vector.

The source detection unit may include: a detection module configured to detect the plurality of human objects from the plurality of consecutive frames for each image source through a bounding box representing a region occupied by the human object using an object detection model, which is the deep learning model, to assign a source identifier; a tracking module configured to assign the same tracking identifier to a bounding box determined to represent the same human object over the plurality of consecutive frames to track the human object; and a selection module configured to extract an image within any one bounding box from a plurality of bounding boxes to which the same tracking identifier is assigned to generate the representative image.

The tracking module may compare areas and positions of a bounding box of a current frame and a bounding box of a previous frame among the plurality of consecutive frames to determine, as the same human object as the bounding box of the previous frame, the bounding box of the current frame in which an area loss representing an area in which the bounding box of the previous frame and the bounding box of the current frame do not overlap is less than a preset first threshold value and a position loss representing a difference between a position of center coordinates of the bounding box of the previous frame and a position of center

coordinates of the bounding box of the current frame is less than a preset second threshold value, and assign the same identifier.

The tracking module may derive the area loss depending on the following Equation.

[Equation]

$$La = 1 - \frac{2\sum_{i}(p_i + q_i) + \varepsilon}{\sum_{i}(p_i^2 + q_i^2) + \varepsilon}$$

Here, La may denote the area loss, i may denote an index of a pixel, q may denote the bounding box of the current frame, p may denote the bounding box of the previous frame, p and q may have a value of 1 when present in the pixel indicated by the index of the pixel, and have a value of 0 when not present, and ε may denote an adjustment constant.

The tracking module may derive the position loss depending on the following Equation.

[Equation]

$$Ld = \| R - S \|_2^2$$

Here, Ld may denote the position loss, R may denote a position vector of the center coordinates of the bounding box of the previous frame, and S may denote a position vector of the center coordinates of the bounding box of the current frame.

The re-identification unit may perform clustering on the target feature vector and the plurality of source feature vectors to derive a plurality of clusters, derive a convex hull from a cluster including the target feature vector among the plurality of clusters, set a center of gravity of a polygon constituting the convex hull as a corresponding cluster center, set, as a

search area, an inner side of a circle having a radius of a distance from a cluster center to a target feature vector based on the cluster center, and calculate a similarity according to a distance between the target feature vector and the source feature vector within the search area.

The re-identification unit may select center vectors of a predetermined number of clusters including the target feature vector and target feature vectors among the plurality of source feature vectors, and perform clustering on the target feature vector and the plurality of source feature vectors while reselecting the center vector until all the source feature vectors are assigned to a cluster with a smallest distance from the center vector of any one cluster to derive the plurality of clusters.

In the conventional case, an enormous number of images were required for re-identification of a human object, but according to the present disclosure, it is possible to reduce resources and improve computation speed by performing re-identification using only one representative image for each image source and human object. Furthermore, according to the present disclosure, it is possible to improve reliability of similarity calculation results by calculating a similarity using not only a similarity distance but also the directionality of a vector space in order to obtain the similarity.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram for describing a configuration of an apparatus for recognizing objects based on artificial intelligence.

FIG. 2 is a diagram for describing a detailed configuration of the apparatus for recognizing objects based on artificial intelligence according to an embodiment of the present disclosure.

FIG. 3 is a diagram for describing a configuration of an object detection model (ODM) for object recognition according to an embodiment of the present disclosure.

FIG. 4 is a flowchart illustrating a method of recognizing objects based on AI according to an embodiment of the present disclosure.

FIGS. 5 and 6 are screen examples for describing the method of recognizing objects based on AI according to the embodiment of the present disclosure.

FIG. 7 is a flowchart for explaining a method of calculating a similarity between a target feature vector and a source feature vector according to an embodiment of the present disclosure.

FIGS. 8 and 9 are diagrams for describing the method of calculating a similarity between a target feature vector and a source feature vector according to the embodiment of the present disclosure.

FIG. 10 is a diagram illustrating a computing device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

Prior to describing in detail the present disclosure, the configurations described in the embodiments and drawings of the present disclosure are merely the most preferable embodiments but do not represent all of the technical spirit of the present disclosure. Thus, the present disclosure should be construed as including all the changes, equivalents, and substitutions included in the spirit and scope of the present disclosure at the time of filing this application.

Hereinafter, embodiments of the present disclosure will be described in detail with reference to the accompanying drawings. At this time, it is noted that like reference numerals denote like elements in appreciating the drawings. Moreover, detailed descriptions related to well-known functions or configurations will be ruled out in order not to unnecessarily obscure the subject matter of the present disclosure. Based on the same reason,

it is to be noted that some components shown in the drawings are exaggerated, omitted or schematically illustrated, and the size of each component does not exactly reflect its real size.

In addition, terms and words used in the present specification and claims to be described below are not to be construed as a general or dictionary meanings but are to be construed as meanings and concepts meeting the technical ideas of the present disclosure based on a principle that the inventors can appropriately define the concepts of terms in order to describe their own inventions in best mode.

First, an apparatus for recognizing objects based on artificial intelligence (AI) according to an embodiment of the present disclosure will be described. FIG. 1 is a diagram for describing a configuration of an apparatus for recognizing objects based on artificial intelligence. FIG. 2 is a diagram for describing a detailed configuration of the apparatus for recognizing objects based on artificial intelligence according to an embodiment of the present disclosure. FIG. 3 is a diagram for describing a configuration of an object detection model (ODM) for object recognition according to an embodiment of the present disclosure.

Referring to FIG. 1, an apparatus 10 for recognizing objects based on artificial intelligence (hereinafter referred to as a “recognition apparatus”) according to an embodiment of the present disclosure includes an image collection unit 100, a source detection unit 200, a source feature detection unit 300, a target selection unit 400, a target feature detection unit 500, and a re-identification unit 600.

The image collection unit 100 is for collecting video including a plurality of consecutive frames from a plurality of image sources. Here, the plurality of image sources may be, for example, a device for capturing a video, such as a closed circuit television (CCTV).

The source detection unit 200 extracts a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality

of consecutive frames of each of a plurality of image sources collected by the image collection unit 100. Referring to FIG. 2, the source detection unit 200 includes a detection module 210, a tracking module 220, and a selection module 230.

The detection module 210 detects a plurality of human objects from a plurality of consecutive frames for each image source through a bounding box (BB) representing a region occupied by the human object using an object detection model (ODM) to assign a source identifier (Source ID). Here, the object detection model (ODM) may be a You Only Look Once (YOLO) series deep learning model.

The tracking module 220 tracks the human object by determining the same human object among the human objects detected by the detection module 210 over the plurality of consecutive frames. To this end, the tracking module 220 assigns the same tracking identifier (Track ID) to the bounding box (BB) determined to represent the same human object over the plurality of consecutive frames. In this way, the tracking module 220 may track the human object in the plurality of consecutive frames.

The selection module 230 generates a representative image by extracting an image within one bounding box from the plurality of bounding boxes to which the same tracking identifier (Track ID) is assigned. That is, a representative image corresponding to the human object tracked through the same tracking identifier (Track ID) is extracted. In this case, the selection module 230 may extract an image of a bounding box having a standard most similar to a preset standard image standard among a plurality of bounding boxes to which the same identifier (Track ID) is assigned to generate the representative image.

Referring back to FIG. 1, the source feature detection unit 300 uses a feature detection model (FDM), which is a deep learning model, to represent a plurality of source features representing features of a plurality of representative images.

When any one human object to be re-identified is selected according to a user's input

from images collected by the image collection unit 100 from a plurality of sources, the target selection unit 400 crops a region including the corresponding human object to extract a query image. The query image is provided to the target feature detection unit 500.

The target feature detection unit 500 derives a target feature vector representing a feature of the query image using the feature extraction model (FDM). In this case, the feature extraction model (FDM) may derive the target feature vector by performing a weight calculation on the query image.

The re-identification unit 600 calculates a similarity between the target feature vector and the plurality of source feature vectors to derive a similarity between the human object to be re-identified in the query image and the human object in the plurality of representative images. Then, the re-identification unit 600 provides representative images in order of the high similarity to the human object to be re-identified.

As described above, the object detection model (ODM) and the feature extraction model (FDM) are deep learning models. The object detection model (ODM) and the feature extraction model (FDM) include a plurality of layers. Each of the plurality of layers is connected with a weight and includes a plurality of operations. Such an operation may exemplify a convolution operation, a pooling operation, or an operation using an activation function. The calculation result of one layer of the object detection model (ODM) and the feature extraction model (FDM) is input to the calculation of the next layer after applying the trained weight. That is, the object detection model (ODM) and the feature extraction model (FDM) calculate outputs by performing a plurality of calculations to which weights between a plurality of layers are applied to inputs. In an embodiment of the present disclosure, a plurality of calculations to which weights between a plurality of layers of the object detection model (ODM) and the feature extraction model (FDM) are applied will be referred to as "weight calculation."

When an image frame is input, the object detection model (ODM) is a deep learning model trained to detect a region occupied by a human object in a frame through a bounding box by performing a weight calculation on the frame. In other words, the object detection model (ODM) represents the region occupied by the human object in the frame through the bounding box.

When the image such as the representative image or the query image is input, the feature extraction model (FDM) is the deep learning model that performs the weight calculation to calculate the source feature vector representing the feature of the representative image or the target feature vector representing the feature of the query image.

Referring to FIG. 3, the feature extraction model (FDM) includes a backbone network (BN), a first pooling layer (PL1: Pooling Layer 1), a second pooling layer (PL2: Pooling Layer 2), a first batch normalization layer (NL1: Batch normalization Layer 1), a second batch normalization layer (NL2: Batch normalization Layer 2), and a fully-connected layer (FL). Here, the backbone network (BN) may be ResNet50. Accordingly, the backbone network (BN) may include a plurality of convolution layers CL.

The backbone network (BN) derives a backbone feature vector through the weight calculation for the representative image or the query image. Then, the first pooling layer (PL1) generates a first pooling feature vector through a pooling (Global Average Pooling) operation on the backbone feature vector, and the second pooling layer (PL2) generates a second pooling feature vector through a pooling (Global Average Pooling) operation on the backbone feature vector. Subsequently, the first batch normalization layer (NL1) generates a first batch normalization feature vector through batch normalization using a moving average and variance of the first pooling feature vector, and the second batch normalization layer (NL2) generates a second batch normalization feature vector through batch normalization using a moving average and variance of the second pooling feature vector. Finally, the fully-

connected layer (FL) derives a source feature vector or a target feature vector by performing a weight calculation using an activation function on the first batch normalization feature vector and the second batch normalization feature vector. Here, the activation function may exemplify sigmoid, hyperbolic tangent (tanh), exponential linear unit (ELU), rectified linear unit (ReLU), leaky ReLU, Maxout, Minout, Softmax, and the like.

The operation of the recognition apparatus 10 including the image collection unit 100, the source detection unit 200, the source feature detection unit 300, the target selection unit 400, the target feature detection unit 500, and the re-identification unit 600 described above will be described in more detail below.

Next, a method of recognizing objects based on AI according to an embodiment of the present disclosure will be described. FIG. 4 is a flowchart illustrating a method of recognizing objects based on AI according to an embodiment of the present disclosure. FIGS. 5 and 6 are screen examples for describing the method of recognizing objects based on AI according to the embodiment of the present disclosure.

Referring to FIG. 4, the image collection unit 100 collects video including a plurality of consecutive frames from a plurality of image sources in step S110. Here, the plurality of image sources may be, for example, a device for capturing a video, such as a closed circuit television (CCTV).

In step S120, the source detection unit 200 extracts a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality of consecutive frames of each of a plurality of image sources. The step S120 will be described in more detail below.

The detection module 210 of the source detection unit 200 detects a plurality of human objects from a plurality of consecutive frames for each image source through the bounding box (BB) representing a region occupied by the human object using the object

detection model (ODM) to assign the source identifier (Source ID). That is, the detection module 210 assigns the same source identifier (Source ID) to all human objects detected in a plurality of frames of images input from the same source. Here, the object detection model (ODM) may be a You Only Look Once (YOLO) series deep learning model. For example, as illustrated in FIG. 4, when a frame F is input, the object detection model (ODM) may detect a human object through a bounding box BB by performing the weight calculation on the frame F. This bounding box BB is defined by center coordinates x and y, a width w, and a height h.

Next, the tracking module 220 of the source detection unit 200 tracks the human object by determining the same human object among the human objects detected by the detection module 210 over the plurality of consecutive frames. To this end, the tracking module 220 assigns the same tracking identifier (Track ID) to the bounding box (BB) determined to represent the same human object over the plurality of consecutive frames. In this way, the tracking module 220 may track the human object in the plurality of consecutive frames. For example, an example of an image including N frames is illustrated in FIG. 5. The tracking module 220 may assign the same tracking identifiers (Track ID1, ID2, and ID3) to the bounding box BB representing the same human object in a first frame F1 to an N-th frame FN.

A more detailed description of this tracking method is as follows. The tracking module 220 compares the areas and positions of the bounding box of the previous frame and the bounding box of the current frame to track the human object. To this end, the tracking module 220 derives an area loss representing the area where the bounding box of the current frame and the bounding box of the previous frame do not overlap according to Equation 1 below.

[Equation 1]

$$La = 1 - \frac{2\sum_{i}(p_i + q_i) + \varepsilon}{\sum_{i}(p_i^2 + q_i^2) + \varepsilon}$$

Here, La denotes the area loss. In addition, i denotes an index of a pixel, q denotes the bounding box of the current frame, p denotes the bounding box of the previous frame. That is, p and q have a value of 1 when present in the pixel indicated by the index of the pixel, and have a value of 0 when not present. ε denotes an adjustment constant.

In addition, the tracking module 220 derives a position loss representing the difference between the position of the center coordinates of the bounding box of the previous frame and the position of the center coordinates of the bounding box of the current frame according to Equation 2 below.

[Equation 2]

$$Ld = \| R - S \|_2^2$$

Here, Ld may denote the position loss. R denotes a position vector of the center coordinates x and y of the bounding box of the previous frame, and S denotes a position vector of the center coordinates x and y of the bounding box of the current frame.

The tracking module 220 compares the bounding box of the previous frame with the bounding box of the current frame to determine, the same human object as the bounding box of the previous frame, as the bounding box of the current frame in which the area loss is less than a preset first threshold value and the position loss is less than a preset second threshold value, thereby assigning the same track identifier (Track ID). In particular, the first

threshold value may be adjusted in proportion to a size of an image, and the second threshold value may be adjusted in proportion to a frame rate of an image.

Next, the selection module 230 of the source detection unit 200 generates a representative image by extracting an image within any one bounding box from the plurality of bounding boxes to which the same tracking identifier (Track ID) is assigned. That is, the representative image corresponding to the human object tracked through the same tracking identifier (Track ID) is extracted. In this case, the selection module 230 may extract an image of a bounding box having a standard most similar to a preset standard image standard among a plurality of bounding boxes to which the same identifier (Track ID) is assigned to generate the representative image. Accordingly, the representative image corresponding to the same human object for each source is detected. In other words, the representative image is detected for each source identifier (Source ID) and tracking identifier (Track ID).

Next, in step S130, the source feature detection unit 300 uses the feature detection model (FDM), which is the deep learning model, to represent a plurality of source features representing features of a plurality of representative images. In this case, the feature extraction model (FDM) may derive the target feature vector by performing the weight calculation on the representative image. This will be described in detail below.

As described above, the feature extraction model (FDM) includes the backbone network (BN), the first pooling layer (PL1), the second pooling layer (PL2), the first batch normalization layer (NL1), the second batch normalization layer (NL2), and the fully-connected layer (FL). First, the backbone network (BN) derives a backbone feature vector through the weight calculation for the representative image. Then, the first pooling layer (PL1) generates the first pooling feature vector through the pooling (Global Average Pooling) operation on the backbone feature vector, and the second pooling layer (PL2) generates the second pooling feature vector through the pooling (Global Average Pooling) operation on the

backbone feature vector. Subsequently, the first batch normalization layer (NL1) generates the first batch normalization feature vector through the batch normalization using the moving average and variance of the first pooling feature vector, and the second batch normalization layer (NL2) generates the second batch normalization feature vector through the batch normalization using the moving average and variance of the second pooling feature vector. Finally, the fully-connected layer (FL) derives the source feature vector by performing the weight calculation using the activation function on the first batch normalization feature vector and the second batch normalization feature vector.

Meanwhile, the query image that is the image of the human object to be re-identified may be selected according to a user's input. In step 140, when any one human object to be re-identified is selected according to a user's input from the images collected by the image collection unit 100 from the plurality of sources, the target selection unit 400 crops the region including the corresponding human object to extract the query image.

In step S150, the target feature detection unit 500 derives the target feature vector representing the feature of the query image using the feature extraction model (FDM). In this case, the feature extraction model (FDM) may derive the target feature vector by performing the weight calculation on the query image. This is the same as the method of deriving the source feature vector. Specifically, the backbone network (BN) derives the backbone feature vector through the weight calculation for the query image. Then, the first pooling layer (PL1) generates the first pooling feature vector through the pooling (Global Average Pooling) operation on the backbone feature vector, and the second pooling layer (PL2) generates the second pooling feature vector through the pooling (Global Average Pooling) operation on the backbone feature vector. Subsequently, the first batch normalization layer (NL1) generates the first batch normalization feature vector through the batch normalization using the moving average and variance of the first pooling feature vector,

and the second batch normalization layer (NL2) generates the second batch normalization feature vector through the batch normalization using the moving average and variance of the second pooling feature vector. Finally, the fully-connected layer (FL) derives the target feature vector by performing the weight calculation using the activation function on the first batch normalization feature vector and the second batch normalization feature vector.

Next, in step S160, the re-identification unit 600 calculates the similarity between the target feature vector and the plurality of source feature vectors to derive the similarity between the human object to be re-identified and the human object of the plurality of representative images, and in step S170, the re-identification unit 600 provides the representative images in order of the high similarity.

Then, the method for calculating a similarity between a target feature vector and a plurality of source feature vectors according to an embodiment of the present disclosure will be described in detail. FIG. 7 is a flowchart for explaining the method of calculating a similarity between a target feature vector and a source feature vector according to the embodiment of the present disclosure. FIGS. 8 and 9 are diagrams for describing the method of calculating a similarity between a target feature vector and a source feature vector according to the embodiment of the present disclosure.

Referring to FIGS. 7 to 9, both the target feature vector and the source feature vector may be referred to as vectors embedded in a vector space (VS) defined through an operation of a feature extraction model. In step S210, the re-identification unit 600 selects a center vector of a predetermined number of clusters including a target feature vector Vtarget and the target feature vector Vtarget among the plurality of source feature vectors on a vector space VS. In this case, the remaining center vectors other than the target feature vector are randomly selected.

Next, the re-identification unit 600 clusters the plurality of source feature vectors

according to a distance from the center vector of the cluster on the vector space VS as illustrated in FIG. 8 in step S220 to form a plurality of clusters (e.g., CL1, CL2, and CL3). That is, the re-identification unit 600 calculates distances between the plurality of source feature vectors and each center vector on the vector space VS as illustrated in FIG. 6, and assigns each of a plurality of source feature vectors to a cluster with a minimum distance from the center vector.

Subsequently, the re-identification unit 600 re-selects the center vector of the cluster for each of the plurality of clusters (e.g., CL1, CL2, and CL3) generated in step S230. For example, the re-identification unit 600 may reselect the source feature vector having a median value among the plurality of source feature vectors allocated to each cluster as the center vector.

Subsequently, the re-identification unit 600 determines whether all source feature vectors are allocated to the cluster having the minimum distance from the center vector of the reselected cluster in step S240. That is, the re-identification unit 600 determines whether to optimize the clustering.

As a result of the determination in step S240, when all the source feature vectors are not assigned to the cluster having the minimum distance from the center vector of the reselected cluster, the re-identification unit 600 repeats steps S220 to S240 described above. That is, the re-identification unit 600 repeats steps S230 to S250 until all source feature vectors are allocated to clusters having the minimum distance from the center vector of the reselected clusters by satisfying Equation 3 below.

[Equation 3]

$$argmin \sum_{i} \sum_{j} f_{ij} \parallel S_i - C_j \parallel^2$$

In Equation 3, i denotes an index of the source feature vector, and Si denotes an i-th

source feature vector. Also, j denotes an index of the cluster, and Cj denotes a center vector of the j-th cluster. fij denotes a flag variable, and when the i-th source feature vector belongs to a j-th cluster, and fij is 1 and otherwise, is 0.

On the other hand, as a result of the determination in step S240, when all the source feature vectors are included in the cluster having the minimum distance from the center vector of the reselected cluster, the re-identification unit 600 determines the current clustering state in step S250 as the optimized state, and completes the clustering.

Then, the re-identification unit 600 derives a convex hull from the cluster including the target feature vector Vtarget in step S260. For example, the cluster including the target feature vector Vtarget is illustrated in FIG. 9. As illustrated, the re-identification unit 600 may configure the convex hull by selecting the plurality of convexes from among the plurality of source feature vectors included in the cluster. Next, the re-identification unit 600 sets a center of gravity of the polygon constituting the convex hull to a center X of the corresponding cluster in step S270. For example, as illustrated in FIG. 9, the re-identification unit 600 may set the center of gravity of the polygon (hexagon in the embodiment of FIG. 9) constituting the convex hull to the center X of the cluster. Although the source feature vector may not exist at the center of gravity of the convex hull, the center X of the cluster is generated by expressing the center of gravity of the convex hull as a vector of the same dimension as the source feature vector.

Subsequently, as illustrated in FIG. 9, in step S280, the re-identification unit 600 sets, as a search area NA, an inner side of a circle whose radius is a distance d from a cluster center X to the target feature vector Vtarget based on the cluster center. Then, the re-identification unit 600 calculates the similarity according to the distance between the target feature vector Vtarget and the source feature vector within the search area NA in step S280. For example, within the cluster CL of FIG. 9, comparing the distances between the target feature vector

Vtarget and the first source feature vector V1, the second source feature vector V2, the third source feature vector V1, and the fourth source feature vector V4, the first source feature vector V1 is closest. However, the first source feature vector V1 is outside the search area NA, and therefore, is excluded from the target. That is, this is because the first source feature vector V1 is located in a direction in which the common feature of the corresponding cluster CL is aligned. On the other hand, when the second source feature vector V2, the third source feature vector V1, and the fourth source feature vector V4 are farther from the target feature vector Vtarget than the first source feature vector V1, but is located in the forward direction of the common feature of the cluster CL, it can be interpreted that they have a feature more similar to the target feature vector Vtarget than the first source feature vector V1. Accordingly, the present disclosure calculates the similarity according to the distance between the target feature vector Vtarget and the source feature vector within the search area NA.

In the conventional case, an enormous number of images were required for re-identification of a human object, but according to the present disclosure, it is possible to reduce resources and improve computation speed by performing re-identification using only one representative image for each image source and human object. Furthermore, according to the present disclosure, it is possible to improve reliability of similarity calculation results by calculating the similarity using not only the similarity distance but also the directionality of a vector space in order to obtain the similarity.

FIG. 10 is a diagram illustrating a computing device according to an embodiment of the present disclosure. A computing device TN100 of FIG. 10 may be a device described in this specification, for example, the recognition apparatus 10.

In the embodiment of FIG. 10, the computing device TN100 may include at least one processor TN110, a transceiver TN120, and a memory TN130. In addition, the computing

apparatus TN100 may further include a storage device TN140, an input interface device TN150, an output interface device TN160, and the like. Components included in the computing apparatus TN100 may communicate with each other by being connected by a bus TN170.

The processor TN110 may execute program commands stored in at least one of the memory TN130 and the storage device TN140. The processor TN110 may mean a central processing unit (CPU), a graphics processing unit (GPU), or a dedicated processor on which methods according to embodiments of the present disclosure are performed. The processor TN110 may be configured to implement procedures, functions, methods, and the like described in relation to embodiments of the present disclosure. The processor TN110 may control each component of the computing apparatus TN100.

Each of the memory TN130 and the storage device TN140 may store various information related to the operation of the processor TN110. Each of the memory TN130 and the storage device TN140 may include at least one of a volatile storage medium and a non-volatile storage medium. For example, the memory TN130 may include at least one of read only memory (ROM) and random access memory (RAM).

The transceiver TN120 may transmit or receive a wired signal or a wireless signal. The transceiver TN120 may perform communication by being connected to a network.

On the other hand, the method according to the embodiment of the present disclosure described above may be implemented in the form of a program readable by various computer means and recorded on a computer-readable recording medium. Here, the recording medium may include program commands, data files, data structures, or the like, alone or in a combination thereof. The program command recorded in the computer-readable recording medium may be especially designed and constituted for the present disclosure or be known to those skilled in a field of computer software. Examples of the recording medium include a

magnetic medium such as a hard disk, a floppy disk, and a magnetic tape; an optical recording medium such as a compact disk read only memory (CD-ROM) and a digital versatile disk (DVD); a magneto-optical medium such as a floptical disk; and a hardware device specially configured to store and perform program commands, such as a ROM, a random access memory (RAM), a flash memory, or the like. Examples of the program commands include a high-level language wire capable of being executed by a computer using an interpreter, or the like, as well as a machine language wire made by a compiler. The abovementioned hardware device may be constituted to be operated as one or more software modules to perform an operation according to the present disclosure, and vice versa.

The present disclosure has been described above using several preferred examples, but these examples are illustrative and not limiting. As such, those skilled in the art to which the present disclosure belongs will understand that various changes and modifications can be made according to the doctrine of equivalents without departing from the spirit of the present disclosure and the scope of rights set forth in the appended claims.

[Detailed Description of Main Elements]

10: Recognition apparatus

100: Image collection unit

200: Source detection unit

210: Detection module

220: Tracking module

230: Selection module

300: Source feature detection unit

400: Target selection unit

500: Target feature detection unit

600: Re-identification unit

5

WHAT IS CLAIMED IS:

1. A method of recognizing objects, comprising:

extracting, by a source detection unit, a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality of consecutive frames of each of a plurality of image sources;

deriving, by a source feature detection unit, a plurality of source feature vectors representing features of the plurality of representative images through a feature extraction model which is a deep learning model;

deriving, by a target feature detection unit, a target feature vector representing a feature of the query image through the feature extraction model when a query image, which is the image of the human object to be re-identified, is input;

deriving, by a re-identification unit, a similarity between the human object to be re-identified and the human object of the plurality of representative images by calculating a similarity between the target feature vector and the plurality of source feature vectors by a re-identification unit; and

providing, by the re-identification unit, representative images in order of the derived high similarity.

2. The method of claim 1, wherein the deriving of the target feature vector includes:

deriving, by a backbone network of the feature extraction model, a backbone feature vector through a weight calculation for the query image;

generating, by a first pooling layer of the feature extraction model, a first pooling feature vector through a pooling operation on the backbone feature vector;

generating, by a second pooling layer of the feature extraction model, a second

pooling feature vector through the pooling operation on the backbone feature vector;

generating, by a first batch normalization layer of the feature extraction model, a first batch normalization feature vector through batch normalization using a moving average and variance of the first pooling feature vector;

generating, by a second batch normalization layer of the feature extraction model, a second batch normalization feature vector through batch normalization using a moving average and variance of the second pooling feature vector; and

performing, by a fully-connected layer of the feature extraction model, a weight calculation using an activation function on the first batch normalization feature vector and the second batch normalization feature vector to derive the target feature vector.

3. The method of claim 1, wherein the detecting of the plurality of representative images includes:

detecting, by a detection module of the source detection unit, the plurality of human objects from the plurality of consecutive frames for each image source through a bounding box representing a region occupied by the human object using an object detection model, which is the deep learning model, to assign a source identifier;

assigning, by a tracking module of the source detection unit, the same tracking identifier to the bounding box determined to represent the same human object over the plurality of consecutive frames to track the human object; and

extracting, by a selection module of the source detection unit, an image within any one bounding box from a plurality of bounding boxes to which the same tracking identifier is assigned to generate the representative image.

4. The method of claim 3, wherein in the tracking of the human object, the tracking

module compares areas and positions of a bounding box of a current frame and a bounding box of a previous frame among the plurality of consecutive frames to determine, as the same human object as the bounding box of the previous frame, the bounding box of the current frame in which an area loss representing an area in which the bounding box of the previous frame and the bounding box of the current frame do not overlap is less than a preset first threshold value and a position loss representing a difference between a position of center coordinates of the bounding box of the previous frame and a position of center coordinates of the bounding box of the current frame is less than a preset second threshold value, and assign the same identifier.

5. The method of claim 4, wherein in the tracking of the human object, the tracking module derives the area loss depending on the following Equation:

[Equation]

$$La = 1 - \frac{2\sum_{i}(p_i + q_i) + \varepsilon}{\sum_{i}(p_i^2 + q_i^2) + \varepsilon}$$

where La denotes the area loss,

i denotes an index of a pixel,

q denotes the bounding box of the current frame,

p denotes the bounding box of the previous frame,

p and q have a value of 1 when present in the pixel indicated by the index of the pixel, and have a value of 0 when not present, and

ε denotes an adjustment constant.

6. The method of claim 4, wherein in the tracking of the human object, the tracking module derives the position loss depending on the following Equation:

[Equation]

$$Ld = \| R - S \|_2^2$$

where Ld denotes the position loss,

R denotes a position vector of the center coordinates of the bounding box of the previous frame, and

S denotes a position vector of the center coordinates of the bounding box of the current frame.

7. The method of claim 1, wherein the deriving of the similarity includes:

performing, by the re-identification unit, clustering on the target feature vector and the plurality of source feature vectors to derive a plurality of clusters;

deriving, by the re-identification unit, a convex hull from a cluster including the target feature vector among the plurality of clusters;

setting, by the re-identification unit, a center of gravity of a polygon constituting the convex hull as a center of a corresponding cluster;

setting, by the re-identification unit, as a search area, an inner side of a circle having a radius of a distance from a cluster center to a target feature vector based on the cluster center; and

calculating, by the re-identification unit, a similarity according to a distance between the target feature vector and the source feature vector within the search area.

8. The method of claim 7, wherein the deriving of the plurality of clusters includes:

selecting, by the re-identification unit, center vectors of a predetermined number of clusters including the target feature vector and target feature vectors among the plurality of source feature vectors; and

performing, by the re-identification unit, clustering on the target feature vector and the plurality of source feature vectors while reselecting the center vector until all the source feature vectors are assigned to a cluster with a smallest distance from the center vector of any one cluster to derive the plurality of clusters.

9. An apparatus for recognizing objects, comprising:

a source detection unit configured to extract a plurality of representative images for each image source and human object by tracking a plurality of human objects in each of a plurality of consecutive frames of each of a plurality of image sources;

a source feature detection unit configured to derive a plurality of source feature vectors representing features of the plurality of representative images through a feature extraction model which is a deep learning model;

a target feature detection unit configured to derive a target feature vector representing a feature of the query image through the feature extraction model when a query image, which is the image of the human object to be re-identified, is input; and

a re-identification unit configured to calculate a similarity between the target feature vector and the plurality of source feature vectors to derive a similarity between the human object to be re-identified and the human object of the plurality of representative images, and provide representative images in order of the derived high similarities.

10. The apparatus of claim 9, wherein the feature extraction model includes:

a backbone network configured to derive a backbone feature vector through a weight calculation for the query image;

a first pooling layer configured to generate a first pooling feature vector through a pooling operation for the backbone feature vector;

a second pooling layer configured to generate a second pooling feature vector through the pooling operation for the backbone feature vector;

a first batch normalization layer configured to generate a first batch normalization feature vector through batch normalization using a moving average and variance of the first pooling feature vector;

a second batch normalization layer configured to generate a second batch normalization feature vector through batch normalization using a moving average and variance of the second pooling feature vector; and

a fully-connected layer configured to perform a weight calculation using an activation function on the first batch normalization feature vector and the second batch normalization feature vector to derive the target feature vector.

11. The apparatus of claim 9, wherein the source detection unit includes:

a detection module configured to detect the plurality of human objects from the plurality of consecutive frames for each image source through a bounding box representing a region occupied by the human object using an object detection model, which is the deep learning model, to assign a source identifier;

a tracking module configured to assign the same tracking identifier to a bounding box determined to represent the same human object over the plurality of consecutive frames to track the human object; and

a selection module configured to extract an image within any one bounding box from a plurality of bounding boxes to which the same tracking identifier is assigned to generate the representative image.

12. The apparatus of claim 11, wherein the tracking module compares areas and positions of a bounding box of a current frame and a bounding box of a previous frame among the plurality of consecutive frames to determine, as the same human object as the bounding box of the previous frame, the bounding box of the current frame in which an area loss representing an area in which the bounding box of the previous frame and the bounding box of the current frame do not overlap is less than a preset first threshold value and a position loss representing a difference between a position of center coordinates of the bounding box of the previous frame and a position of center coordinates of the bounding box of the current frame is less than a preset second threshold value, and assign the same identifier.

13. The method of claim 12, wherein the tracking module derives the area loss depending on the following Equation:

[Equation]

$$La = 1 - \frac{2\sum_{i}(p_i + q_i) + \varepsilon}{\sum_{i}(p_i^2 + q_i^2) + \varepsilon}$$

where La denotes the area loss,

i denotes an index of a pixel,

q denotes the bounding box of the current frame,

p denotes the bounding box of the previous frame,

p and q have a value of 1 when present in the pixel indicated by the index of the pixel, and have a value of 0 when not present, and

ε denotes an adjustment constant.

14. The apparatus of claim 12, wherein the tracking module derives the position loss depending on the following Equation:

[Equation]

$$Ld = \| R - S \|_2^2$$

where Ld denotes the position loss,

R denotes a position vector of the center coordinates of the bounding box of the previous frame, and

S denotes a position vector of the center coordinates of the bounding box of the current frame.

15. The apparatus of claim 9, wherein the re-identification unit performs clustering on the target feature vector and the plurality of source feature vectors to derive a plurality of clusters,

derives a convex hull from a cluster including the target feature vector among the plurality of clusters,

sets a center of gravity of a polygon constituting the convex hull as a center of a corresponding cluster,

sets, as a search area, an inner side of a circle having a radius of a distance from a cluster center to a target feature vector based on the cluster center, and

calculates a similarity according to a distance between the target feature vector and

the source feature vector within the search area.

16. The apparatus of claim 15, wherein the re-identification unit selects center vectors of a predetermined number of clusters including the target feature vector and target feature vectors among the plurality of source feature vectors, and

performs clustering on the target feature vector and the plurality of source feature vectors while reselecting the center vector until all the source feature vectors are assigned to a cluster with a smallest distance from the center vector of any one cluster to derive the plurality of clusters.