

**METHOD AND APPARATUS FOR DETECTING CROWD BEHAVIOR,
ELECTRONIC DEVICE, STORAGE MEDIUM AND COMPUTER PROGRAM
PRODUCT**

CROSS REFERENCE TO RELATED APPLICATIONS

[01] This application claims priority to Chinese Patent Application No. 2021101062857 entitled “METHOD AND APPARATUS FOR DETECTING CROWD BEHAVIOR, ELECTRONIC DEVICE, AND STORAGE MEDIUM” filed on January 26, 2021, the entire content of which is incorporated herein by reference.

TECHNICAL FIELD

[02] The present disclosure relates to computer technologies, and in particular to a method and an apparatus for detecting a crowd behavior, an electronic device, a storage medium and a computer program product.

BACKGROUND

[03] Along with progress of urbanization, crowds become more and more concentrated. Therefore, it is very important for the safety of pedestrians to identify whether an abnormal behavior occurs in a crowd and what is the abnormal behavior. If an abnormal behavior of a crowd can be accurately identified and stopped, a probability of occurrence of a dangerous event can be reduced.

[04] For example, in a security protection scenario, a target image sequence (video sequence) including pedestrians can be captured by an image capture device (e.g. monitoring device). If a pedestrian behavior being occurring in the target image sequence belongs to an abnormal behavior such as pedestrians gathering and pedestrians stranding or the like, crowd evacuation can be carried out immediately to avoid occurrence of stampedes or vicious crowd incidents. Thus, it is urgent to propose a method of detecting a crowd behavior in a target image sequence.

SUMMARY

[05] In view of this, the present disclosure at least provides a method of detecting a crowd behavior. The method includes: performing object tracking for at least one object in a target image sequence including objects to determine position change information of each of the at least one object in the target image sequence; based on the position change information obtained in the target image sequence, performing graph convolution processing; and based on extracted features obtained by the graph convolution processing, determining a crowd behavior corresponding to the objects in the target image sequence.

[06] In some illustrated embodiments, performing object tracking for the at least one object in the target image sequence including the objects to determine the position change information of each of the at least one object in the target image sequence includes: performing image processing for each image in the target image sequence to determine position information of each of the at least one object in a corresponding image; performing object tracking for each of the at least one object so as to determine the position change information of each of the at least one object in the target image sequence based on a tracking result and the position information.

[07] In some illustrated embodiments, performing object tracking for each of the at least one object so as to determine the position change information of each of the at least one object in the target image sequence based on the tracking result and the position information includes: by using a Kalman filter algorithm or an object detection model, performing object tracking for each of the at least one object; based on the position information of a same tracked object in corresponding images, determining the position change information of each of the at least one object.

[08] In some illustrated embodiments, based on the position change information obtained in the target image sequence, performing graph convolution processing to determine a crowd behavior corresponding to the objects in the target image sequence includes: based on position information of objects in at least one image in the target image sequence represented by the position change information and an association relationship between the objects in the at least one image, respectively performing spatial graph convolution processing for each of the at least one image to obtain a respective graph feature corresponding to each of the at least one image; performing time-domain convolution processing for the graph feature corresponding to each of the at least one image, and based on extracted features obtained by the time-domain convolution processing, determining the crowd behavior corresponding to the objects in the target image sequence; wherein the crowd behavior includes at least one of the followings: pedestrians gathering, pedestrians dispersion, pedestrians stranding, or pedestrians reversing.

[09] In some illustrated embodiments, based on the position information of objects in the at least one image in the target image sequence represented by the position change information and the association relationship between the objects in the at least one image, performing the spatial graph convolution processing for each of the at least one image respectively to obtain the graph feature corresponding to each of the at least one image

includes: based on the association relationship between the objects in the at least one image, determining a respective adjacency matrix corresponding to each of the at least one image; based on the position information of the objects in the at least one image, determining a respective feature matrix corresponding to each of the at least one image; for each of the at least one image, based on the adjacency matrix corresponding to the image and the feature matrix corresponding to the image, performing the spatial graph convolution processing to obtain a graph feature corresponding to the image.

[10] In some illustrated embodiments, before performing the step of performing graph convolution processing based on the position change information obtained in the target image sequence to obtain the extracted features corresponding to the target image sequence, the method further includes: determining an association relationship between any two objects included in at least one image in the target image sequence.

[11] In some illustrated embodiments, determining the association relationship between any two objects included in at least one image in the target image sequence includes: for each of the at least one object comprised in the at least one image, extracting an image feature corresponding to a region where the object is located in the image, wherein the image feature represents image information of a position where the object is located; based on the image feature corresponding to the at least one object, determining a similarity between any two objects in the at least one object; determining two objects with a similarity less than a first preset threshold as two objects having an association relationship.

[12] In some illustrated embodiments, determining the association relationship between any two objects included in at least one image in the target image sequence includes: respectively performing image processing for each of the at least one image to determine position information of the at least one object in the at least one image; based on the position information corresponding to the at least one object, determining a distance between any two objects in the at least one object; based on the distance, determining the association relationship between the any two objects included in the at least one image.

[13] In some illustrated embodiments, based on the distance, determining the association relationship between any two objects included in the at least one image includes: mapping the determined distance between any two objects to an interval formed by a third preset threshold and a fourth preset threshold; determining the mapped distance between the any two objects as an association weight between the any two objects; by using the

association weight between the any two objects, indicating the association relationship between the any two objects.

[14] In some illustrated embodiments, the graph convolution processing is implemented by using a graph convolution classification model; where, the graph convolution classification model is trained by: generating training samples, where each of the training samples has position change information of objects and labeling information of a crowd behavior based on the position change information of the objects; based on the position change information and the labeling information of the crowd behavior, training a preset graph convolution model to obtain the graph convolution classification model.

[15] In some illustrated embodiments, generating one of the training samples includes: based on a motion simulation platform, setting motion modes corresponding to virtual objects; based on the motion modes, determining position change information corresponding to at least one of the virtual objects; determining a crowd behavior represented by the position change information corresponding to the at least one of the virtual objects; based on the position change information and the crowd behavior represented by the position change information, generating the training sample.

[16] The present disclosure further provides an apparatus for detecting a crowd behavior. The apparatus includes: a position change information determining module, configured to, based on an object tracking result of at least one object in a target image sequence including objects, determine position change information of each of the at least one object in the target image sequence; a crowd behavior detecting module, configured to, based on the position change information obtained in the target image sequence, perform graph convolution processing, and based on extracted features obtained by the graph convolution processing, determine a crowd behavior corresponding to the objects in the target image sequence.

[17] The present disclosure further provides an electronic device. The electronic device includes a processor and a memory storing instructions executable by the processor, where the processor is configured to invoke the executable instructions stored in the memory to implement the above method of detecting a crowd behavior.

[18] The present disclosure further provides a computer readable storage medium, storing computer programs, where the computer programs are executed to implement the above method of detecting a crowd behavior.

[19] The present disclosure further provides a computer program product, where the computer program product is run on a computer to cause the computer to implement the above method of detecting a crowd behavior.

[20] In the present disclosure, by performing object tracking for objects occurring in a target image sequence, position change information of the objects in the target image sequence is determined. Then, based on the position change information, graph convolution processing is performed to obtain extracted features corresponding to the target image sequence, and based on the extracted features, a crowd behavior corresponding to the objects in the target image sequence is determined. In this way, based on the principle of the graph convolution, extracted features used to detecting a crowd behavior can be determined from the target image sequence, thus accurately detecting a crowd behavior represented by the objects in the target image sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

[21] FIG. 1 is a flowchart illustrating a classification method of a target image sequence according to an embodiment of the present disclosure.

[22] FIG. 2 is a schematic diagram illustrating a process of detecting a crowd behavior according to an embodiment of the present disclosure.

[23] FIG. 3 is flowchart illustrating a method of determining an association relationship of objects in an image according to an embodiment of the present disclosure.

[24] FIG. 4 is a schematic diagram illustrating a process of performing graph convolution processing according to an embodiment of the present disclosure.

[25] FIG. 5 is a schematic diagram illustrating a classification process according to an embodiment of the present disclosure.

[26] FIG. 6 is a schematic diagram illustrating a video sequence classification process according to an embodiment of the present disclosure.

[27] FIG. 7 is a flowchart illustrating a method of training a model according to an embodiment of the present disclosure.

[28] FIG. 8 is a structural schematic diagram illustrating an apparatus for detecting a crowd behavior according to an embodiment of the present disclosure.

[29] FIG. 9 is a schematic diagram illustrating a hardware structure of an electronic device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[30] Exemplary embodiments will be described in detail herein, with the illustrations thereof represented in the drawings. When the following descriptions involve the drawings, like numerals in different drawings refer to like or similar elements unless otherwise indicated. The embodiments described in the following examples do not represent all embodiments consistent with the present disclosure. Rather, they are merely examples of apparatuses and methods consistent with some aspects of the present disclosure as detailed in the appended claims.

[31] Terms used herein are used to only describe a particular embodiment rather than limit the present disclosure. The singular forms such as “a”, “said”, and “the” used in the present disclosure and the appended claims are also intended to include multiple, unless the context clearly indicates otherwise. It is also to be understood that the term “and/or” as used herein refers to and includes any or all possible combinations of one or more associated listed items. It should be understood that, depending on the context, the term “if” as used herein may be interpreted as “when” or “upon” or “in response to determining”.

[32] The present disclosure aims to provide a method of detecting a crowd behavior (hereinafter referred to as detection method). In the detection method, based on the graph convolution principle and position change information corresponding to each object in a target image sequence, extracted features to determine a crowd behavior corresponding to objects in the target image sequence are obtained from the target image sequence. Next, the method may continue to perform classification based on the extracted features so as to determine a crowd behavior corresponding to the objects in the target image sequence.

[33] For example, in a special scenario, the target image sequence may be a video sequence captured by a monitor, and the objects may be pedestrians occurred in the target image sequence. The type of the crowd behavior may include pedestrians gathering, pedestrians stranding and pedestrians dispersion and the like. In this method, based on graph convolution principle and position change information of the pedestrians in a video, extracted features helpful to determining a crowd behavior may be determined. Next, based on the extracted features, classification is performed to determine a crowd behavior occurring in the video sequence, and then make corresponding arrangement based on the crowd behavior, thus reducing a probability of occurrence of dangerous incidents.

[34] FIG. 1 is a flowchart illustrating a classification method of a target image sequence according to an embodiment of the present disclosure.

[35] As shown in FIG. 1, the method may include the following steps.

[36] At step S102, object tracking is performed for at least one object in a target image sequence including objects to determine position change information of each of the at least one object in the target image sequence.

[37] At step S104, based on the position change information obtained in the target image sequence, graph convolution processing is performed; and based on extracted features obtained by the graph convolution processing, a crowd behavior corresponding to the objects in the target image sequence is determined.

[38] The classification method may be applied to an electronic device. The electronic device may perform the classification method by using a software corresponding to the classification method. The electronic device may be a laptop computer, a computer, a server, a smart phone, a PAD (Packet Assembler And Disassembler) terminal or the like. The present disclosure does not make limitations to the specific type of the electronic device.

[39] It may be understood that the classification method can be performed by a terminal device or a server device separately or in cooperation.

[40] For example, the classification method may be integrated into a client. A terminal device provided with the client may, after receiving a classification request, provide computing power through its own hardware environment to execute the classification method.

[41] For another example, the classification method may be integrated into a system platform. A server device provided with the system platform may, after receiving a classification request, provide computing power through its own hardware environment to execute the classification method.

[42] For another example, the classification method is to perform two tasks: obtaining the target image sequence and classifying the target image sequence, where the obtaining task may be integrated into a client provided on a terminal device and the classification task may be integrated into a server provided on a server device. After obtaining the target image sequence, the terminal device may send a classification request to the server device. After receiving the classification request, the server device may perform the classification method for the target image sequence in response to the request. The following descriptions are made with an electronic device (referred to as device hereinafter) as an execution subject.

[43] FIG. 2 is a schematic diagram illustrating a process of detecting a crowd behavior according to one or more embodiments of the present disclosure.

[44] Before steps shown in FIG. 2 is performed, a target image sequence may be obtained. The target image sequence refers to an image sequence including pedestrian objects to be detected for crowd behavior. The target image sequence may include a plurality of images.

[45] In some examples, the target image sequence may include a video sequence or a sequence of discrete image frames. The video sequence may include N continuous image frames including objects, where N is a positive integer.

[46] In some examples, when obtaining the target image sequence, the device may interact with a user to input the target image sequence. For example, the device may provide a window for a user to input a to-be-processed target image sequence by using an interface provided on itself, such that the user may input the target image sequence. The user may input the target image sequence through the window.

[47] In some examples, the device may be also coupled with an image capture device (e.g. a video monitor) deployed on site, so as to obtain the target image sequence by the image capture device.

[48] After the target image sequence is obtained, the step S102 may be implemented to perform object tracking for at least one object in the target image sequence including objects to determine position change information of each object in the target image sequence. The object tracking specifically refers to performing tracking for a same object in images. When the object tracking is performed, the object tracking can be completed by determining a same object in the images. For example, in a security protection scenario, the object tracking is pedestrian tracking. When the pedestrian tracking is performed, the pedestrian tracking can be completed by determining a same pedestrian in the images.

[49] The position change information specifically refers to motion trajectory information of an object in the target image sequence. For example, pedestrian tracking can be performed for a pedestrian to determine the position information of the same pedestrian in images, so as to determine a motion trajectory of the pedestrian in the image sequence. It may be understood that the position change information may represent object position information of an object in each image and time-domain information. The object position information may represent an object coordinate. The time-domain information may represent time information corresponding to an object at each position.

[50] With continuous reference to FIG. 2, in the present disclosure, the obtained target image sequence may be input into an object tracking unit to perform the step S102.

[51] The object tracking unit may, by using device-executable instructions, perform step S1022 of performing position prediction processing for each image based on an object position prediction model to determine position information of each object in each image. In this step, by using the object position prediction model, position prediction processing may be performed for each image to determine the position information of the object in each image. The object position prediction model includes a model trained based on training samples labeled with the object position information.

[52] It may be understood that, the object position prediction model may be a neural network model constructed based on a deep convolutional network. Before position prediction is performed using the model, supervised training may be performed for the position prediction model by using the training samples labeled with the object position information until the model converges.

[53] After the position information is determined, the object tracking unit may perform step S1024 of performing object tracking for the object based on the position information to determine the position change information of the object in the target image sequence.

[54] The present disclosure does not make specific limitations to the method of performing object tracking, and two illustrative object tracking methods are provided below.

[55] In a first method, a Kalman filter algorithm may be used in step S1024 to perform object tracking for each object to determine the position change information of each object.

[56] In some examples, according to a capture sequence of the images, starting from the first captured image, two adjacent images are taken as two current images sequentially and the following steps are performed: by using the Kalman filter algorithm, determining the position information corresponding to each object included in the two current images; by using the Hungarian matching algorithm (bipartite graph matching algorithm), respectively comparing the position information corresponding to each object in a first image of the two current images with the position information corresponding to each object in a second image of the two current images.

[57] When the comparison operation is performed, a distance between the position information corresponding to each object included in the first image and the position information corresponding to each object included in the second image may be calculated. If the calculated distance is less than a preset standard threshold, it is determined that the two pieces of position information corresponding to the distance are two pieces of matched position formation.

[58] After the comparison operation is completed, two objects corresponding to the two pieces of matched position information are determined as a same object occurred in the two current images, thus achieving object tracking for the object.

[59] After the above step is completed for all adjacent images, the position change information of the object may be determined based on the position information of the same tracked object in each image.

[60] In the method, the same object occurred in each image can be determined so as to achieving tracking for the same object in each image. After the object tracking for the object is achieved, the position change information of the object in the target image sequence can be determined based on the position information of the object in each image.

[61] In a second method, a same object occurred in each image may be determined based on an object detection model in step S1024, so as to perform object tracking for each object.

[62] The object detection model includes a model constructed based on deep learning network. For example, the object detection model may specifically be a pre-trained semantic detection model (e.g. fast-rcnn, faster-rcnn, mask-rcnn or the like). By using the detection model, an object feature corresponding to a pedestrian object included in an image can be detected. In some examples, the object feature may be a human face feature. After the object feature included in each image is detected, similarity calculation may be performed for object features included in two different images, and the objects with a similarity greater than or equal to a second standard threshold are determined as a same object.

[63] For example, in a security protection scenario, the object may be a pedestrian. At this time, a human face included in each image may be detected by using the object detection model. After the human face in each image is detected, similarity calculation may be performed for human face features included in two different images and the human faces with a similarity greater than or equal to the second standard threshold are determined as a same human face. By determining the same human face, it is determined that a same pedestrian is occurred in the two images.

[64] After the same object in each image is determined, the position change information of the same object may be determined based on the position information of the same tracked object.

[65] In some examples, after the position change information of the object is determined, the position change information corresponding to each object may be stored in the form of three-dimensional matrix ($T*H*W$). The three-dimensional matrix may have a number of

channels equal to a number of images included in the target image sequence. The element of the three-dimensional matrix may be a position coordinate of an object in an image corresponding to a channel sequence number. It may be understood that the three-dimensional matrix can be determined as a feature matrix corresponding to the target image sequence at this time.

[66] It may be understood that the position change information has time-domain features, which can indicate change of the position coordinate of an object during a motion process within a time-domain range shown in the target image sequence. Based on the position change information corresponding to each object in the target image sequence, motion characteristic of each object can be determined, that is, whether the objects gradually gather or gradually disperse can be determined. Therefore, it is feasible to detect a crowd behavior based on the position change information.

[67] With continuous reference to FIG. 2, after the position change information is determined, step S104 may be performed to, based on the position change information obtained in the target image sequence, perform graph convolution processing, and based on extracted features obtained by the graph convolution processing, determine a crowd behavior corresponding to the objects in the target image sequence.

[68] Firstly, step S1042 may be carried out to perform graph convolution processing based on the position change information obtained in the target image sequence to obtain extracted features corresponding to the target image sequence.

[69] The extracted feature specifically includes a feature matrix or a feature vector determined by graph convolution processing (including spatial graph convolution processing and time-domain graph convolution processing). It may be understood that the extracted features are determined based on the position change information of pedestrian objects in the target image sequence, and thus, the extracted features are beneficial to determining a crowd behavior.

[70] In some examples, before step S1042 is performed, an association relationship between objects in each image of the target image sequence may be determined.

[71] It may be understood that the association relationships determined based on different rules have different meanings. For example, the association relationship determined based on a similarity between image features corresponding to regions where two objects are located in the images may represent an association degree of the two objects from the perspective of similarity. For another example, the association relationship

determined based on a distance between two objects may represent an association degree of the two objects from the perspective of distance.

[72] In some examples, for each image in the target image sequence, the image feature corresponding to a region where an object included in the image is located in the image may be extracted. The image feature represents image information of a position where the object is located. If two objects have similar image features, it indicates that the two objects are similar in position, that is, the two objects have positions close to each other and thus have an association relationship.

[73] Afterwards, based on the image feature corresponding to each object, a similarity between any two objects may be determined.

[74] In some examples, two objects corresponding to a similarity greater than or equal to a first preset threshold are determined as two objects having an association relationship. The first preset threshold may be set based on experiences. In the present disclosure, no specific limitation is made to the first preset threshold.

[75] It is noted that the present disclosure does not make limitations to the method of calculating a similarity. For example, the method of calculating a similarity may be a method such as Euclidean distance, cosine distance and Mahalanobis distance. In some examples, in order to increase the classification accuracy for the target image sequence, the association relationship between objects may be determined based on a distance between objects.

[76] FIG. 3 is a schematic flowchart illustrating a method of determining an association relationship according to one or more embodiments of the present disclosure. As shown in FIG. 3, step S302 may be carried out to perform image processing for each image to determine position information of the objects in each image. Afterwards, step S304 may be carried out to determine a distance between any two objects in the objects based on the position information corresponding to each object.

[77] After the distance between any two objects is determined, an association relationship between any two objects included in each image may be determined based on the distance. In some examples, when the association relationship between any two objects included in the images is determined based on the distance, two objects corresponding to a distance less than a second preset threshold are determined as the two objects having an association relationship. The second preset threshold may be determined based on experiences. The present disclosure does not make specific limitations to the second preset threshold.

[78] In some examples, if two objects are determined to have an association relationship, an association weight between the two objects is set to 1, and otherwise, set to 0.

[79] Since the association relationship is determined based on distance between objects, a spatial-temporal graph determined based on the association relationship may indicate a distance relationship between objects, and the extracted features determined by performing graph convolution processing for the spatial-temporary graph may also include distance information between objects. Therefore, when classification is performed for a crowd behavior in the target image sequence based on the extracted features, the classification accuracy of pedestrians gathering, pedestrians dispersion or pedestrians stranding can be improved.

[80] In some examples, in order to further improve the classification accuracy, an association weight between two objects can be determined based on a true distance between the two objects.

[81] Specifically, the determined distance between any two objects is mapped to an interval formed by a third preset threshold and a fourth preset threshold, where the third preset threshold and the fourth preset threshold are empirical thresholds. In some examples, the third preset threshold is 0 and the fourth preset threshold is 1.

[82] After the mapping is completed, the mapped distance between any two objects is determined as an association weight between the any two objects, which is used to indicate an association relationship between the any two objects.

[83] Since the association relationship between two objects is determined based on the true distance between the two objects in the above examples, the spatial-temporary graph can indicate distance information closer to the reality, thus further improving the classification accuracy.

[84] After the association relationship between any two objects included in each image in the target image sequence is determined, step S104 may be continued.

[85] With continuous reference to FIG. 2, the step S1042 may be performed by using a graph convolution model.

[86] The graph convolution model may be a model constructed based on a spatial-temporal graph convolution processing network. The spatial-temporal graph convolution processing network at least includes a spatial graph convolution network (GCN) for performing spatial graph convolution processing for each image, and a time-domain

convolution network (TCN) for performing time-domain convolution processing for the graph feature corresponding to each image.

[87] FIG. 4 is a schematic diagram illustrating a graph convolution processing process according to one or more embodiments of the present disclosure.

[88] As shown in FIG. 4, the position change information may be input into the GCN included in the graph convolution model in the step S1042 to perform step S402 of, based on position information of the objects in each image in the target image sequence represented by the position change information and an association relationship between objects in each image, respectively performing spatial graph convolution processing for each image to obtain a respective graph feature corresponding to each image.

[89] In this step, based on the association relationship between objects in the images, a respective adjacency matrix corresponding to each image may be determined. In some examples, a topological diagram corresponding to each image may be firstly generated. In practical applications, for each image, each object in the image may be taken as a vertex V of the topological diagram, and a side E is determined based on the association relationship between objects in the image, so as to obtain a topological diagram corresponding to the image.

[90] After the topological diagram corresponding to each image is generated, for each image, an adjacency matrix A corresponding to the image may be determined based on the topological diagram corresponding to the image, and based on the object position information, a feature matrix X^0 corresponding to the image may be determined.

[91] For each image, after the adjacency matrix for the image and the feature matrix for the image are determined, the spatial graph convolution processing can be completed based on the adjacency matrix for the image and the feature matrix for the image to obtain a graph feature corresponding to the image.

[92] It is noted that no specific limitations are made to the graph convolution formula in the present disclosure. In some examples, $X^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l)}\theta)$ may be adopted. where $\tilde{A} = A + I$, which is used for increasing the features of self-loop, $\tilde{D}$ indicates a diagonal matrix of $\tilde{A}$, θ indicates a network parameter of the graph convolution network (the specific training process is shown in the subsequent contents of the present disclosure, and will not be repeated herein), $X^{(l)}$ indicates an input of the $(l+1)$ -th hidden layer in the GCN, and $X^{(l+1)}$ indicates an output after operation of the $(l+1)$ -th hidden layer.

[93] After the graph feature corresponding to each image is obtained, the graph feature may be input into the TCN included in the graph convolution model to perform step S404 of performing time-domain convolution processing for the graph feature corresponding to each image of the target image sequence to obtain extracted features corresponding to the target image sequence.

[94] In this step, the graph features corresponding to images in the target image sequence may be sorted based on time-domain information represented by the position change information, and then based on a preset one-dimensional convolution kernel, one-dimensional convolution processing is performed for the sorted graph features corresponding to images in the target image sequence to obtain extracted features corresponding to the target image sequence.

[95] With continuous reference to FIG. 2, after the extracted features corresponding to the target image sequence are obtained, step S1044 may be carried out to determine a crowd behavior corresponding to objects in the target image sequence based on the extracted features.

[96] In this step, the extracted features may be input into a pre-trained multi-classifier for classification so as to obtain the crowd behavior.

[97] FIG. 5 is a schematic diagram illustrating a classification process according to an embodiment of the present disclosure.

[98] As shown in FIG. 5, the multi-classifier includes a down-sampling unit and a fully-connected layer. The down-sampling unit may be used to process the extracted features to obtain the corresponding feature vectors. For example, the down-sampling unit may be an average pooling unit. The fully-connected layer is used to perform classification based on the feature vectors to obtain a confidence score corresponding to each preset classification type.

[99] With continuous reference to FIG. 5, in step S1044, the extracted features may be input into the down-sampling unit to perform step S502 of performing average pooling for the extracted features to obtain corresponding feature vectors. After the feature vectors are obtained, the feature vectors may be input into the fully-connected layer to perform step S504 of performing fully-connected processing for the feature vectors to obtain a confidence score corresponding to each preset classification type.

[100] After each confidence score is obtained, a crowd behavior type corresponding to a maximum confidence score may be determined as a crowd behavior corresponding to the

objects in the target image sequence. The crowd behavior includes at least one of the followings: pedestrians gathering, pedestrians dispersion, pedestrians stranding and pedestrians reversing.

[101] In the above method, object tracking is performed for objects occurred in a target image sequence to determine position change information of the objects in the target image sequence. Then, based on the position change information, graph convolution processing is performed to obtain extracted features corresponding to the target image sequence, and based on the extracted features, a crowd behavior corresponding to objects in the target image sequence is determined. In this way, based on the principle of the graph convolution, the extracted features beneficial to detecting a crowd behavior are determined from the target image sequence, thus accurately detecting a crowd behavior represented by the target image sequence.

[102] An embodiment is described below in combination with a security protection scenario. The security protection scenario is usually provided with a monitoring device. The monitoring device is usually capable of capturing a video sequence. It may be understood that, the video sequence captured by the monitoring device is actually classified in the security protection scenario. FIG. 6 is a schematic diagram of a video sequence classification process according to an embodiment of the present disclosure.

[103] After the target video sequence is obtained, a coordinate determining unit may be used to perform step S602 of performing image processing for each image in the target video sequence to determine position information of pedestrians occurred in the video in each image.

[104] After the position information is determined, a pedestrian tracking unit may be used to perform step S604 of performing object tracking for the pedestrians based on the position information to determine position change information of the pedestrians in the target image sequence.

[105] After the position change information is determined, a graph convolution model included in the graph convolution classification model may be used to perform step S606 of performing graph convolution processing based on the position change information to obtain extracted features corresponding to the target image sequence.

[106] The graph convolution classification model specifically may be a classification model constructed based on graph convolution model and multi-classification model. With the graph convolution classification model, on one hand, graph convolution operation may

be performed for a spatial-temporary graph to determine extracted features corresponding to the spatial-temporal graph; on the other hand, classification is performed for the target image sequence based on the extracted features to determine a classification type of the sequence.

[107] After the extracted features are determined, the multi-classification model included in the graph convolution classification model may be used to perform step S608 of determining a crowd behavior corresponding to the objects in the target image sequence based on the extracted features.

[108] In the above solution, based on graph convolution principle and the position change information of the pedestrians in the video, extracted features reflecting distance change information of each pedestrian in the video sequence are firstly determined. Then, based on the extracted features, the classification type of the video sequence is determined so as to determine a pedestrian behavior occurring in the video sequence and take corresponding measures based on the pedestrian behavior, thus reducing the probability of occurrence of dangerous incidents.

[109] The above introduction is made to the classification solution of the image sequence shown in the present disclosure and the following descriptions are made to the method of training a graph convolution classification model.

[110] The graph convolution classification model may be used to achieve the graph convolution processing.

[111] In some examples, the graph convolution classification model may include a graph convolution model and a multi-classification model. The graph convolution model may use the position change information of each object in the target image sequence as input to perform graph convolution processing to obtain extracted features corresponding to the target image sequence. The multi-classification model may perform classification for the extracted features as input to obtain a crowd behavior represented by the target image sequence.

[112] It may be understood that, the process for training the graph convolution classification model actually is a process for training model parameters included in the graph convolution model and the multi-classification model.

[113] The present disclosure provides a model training method. The method trains the graph convolution classification model by constructing virtual training samples, such that model training can be achieved in a case of lack of true samples.

[114] FIG. 7 is a flowchart illustrating a model training method according to the present disclosure. As shown in FIG. 7, the training method includes: at step S702, generating training samples, where each of the training sample has position change information of objects and labeling information of a crowd behavior based on the position change information of the objects.

[115] In this step, step S7022 may be performed firstly to set a motion mode corresponding to an object occurred in the video based on a motion simulation platform.

[116] The motion simulation platform specifically may be any platform carrying out motion simulation. In some examples, the motion simulation platform may be a game development platform.

[117] The motion mode may include a velocity and a motion direction. In the motion mode, on one hand, a coordinate of each object in each image included in the video may be determined to determine position change information of each object in the video; on the other hand, a crowd behavior represented by the video can be obtained. For example, in a security protection scenario, when the motion mode of each pedestrian is towards a same direction, a crowd behavior represented by the video is determined as pedestrians gathering, and otherwise, determined as pedestrians dispersion.

[118] After the motion mode of each object is determined, step S7024 may be carried out to, based on the motion mode, determine the position change information corresponding to each object, and determine the crowd behavior represented by the position change information corresponding to each object, where the crowd behavior may include pedestrians gathering, pedestrians dispersion and pedestrians stranding and the like.

[119] After the position change information and the crowd behavior represented by the video are determined, step S7026 may be carried out to, based on the position change information and the crowd behavior represented by the position change information, generate the training sample.

[120] In this step, encoding may be performed for the position change information and the classification type based on one-hot encoding or the like, so as to obtain a plurality of training samples. The present disclosure does not make limitation to the specific manner of the encoding.

[121] After the training samples are obtained, step S704 may be carried out to, based on preset loss information and the training samples, train the graph convolution classification model until the model converges.

[122] The preset loss information may be loss information set based on experiences.

[123] During model training, hyper-parameters such as learning rate and training cycle number are firstly specified. After the hyper-parameters are determined, supervised training may be performed for the graph convolution classification model (referred to as the model) based on the training samples.

[124] During a supervised training process, forward propagation may be performed to obtain a calculation result output by the model. After the calculation result output by the model is obtained, an error between a true classification type and the calculation result may be evaluated based on the preset loss information. After the error is obtained, a descent gradient is determined based on stochastic gradient descent method. After the descent gradient is determined, the model parameter corresponding to the model may be updated based on backward propagation.

[125] Next, the above process is repeated until the model converges. It is noted that a condition for the model convergence may be that a preset training number is reached, or a change amount of error obtained after M number of continuous forward propagations is less than a given threshold or the like, which is not limited in the present disclosure.

[126] In the above training method, the graph convolution classification model is trained using the training samples without relying on the true training samples in a training process.

[127] In some examples, joint training may be further performed for the object position prediction model to determine an object position, an object tracking model to perform object tracking and the graph convolution classification model to perform graph convolution processing and classification.

[128] In some examples, a video representing pedestrian gathering and pedestrian dispersion and the like may be constructed by using a motion simulation platform and crowd behavior labeling is performed for the constructed video to obtain training samples.

[129] After the training samples are obtained, the training samples may be input into the object position prediction model to obtain a first calculation result. Next, the first calculation result is input the object tracking model to obtain a second calculation result. Next, the second calculation result is input into the graph convolution classification model to obtain a detection result for a crowd behavior represented by the video.

[130] After the detection result is obtained, update for the parameters of each model can be completed by using backward propagation based on the labeling information corresponding to the virtual training samples.

[131] In the above examples, joint training for various models can be achieved so as to improve the training efficiency.

[132] Corresponding to any one of the above embodiments, the present disclosure further provides an apparatus for detecting a crowd behavior.

[133] FIG. 8 is a structural schematic diagram illustrating an apparatus for detecting a crowd behavior according to the present disclosure.

[134] As shown in FIG. 8, the apparatus 80 includes: a position change information determining module 81, configured to, based on an object tracking result of at least one object in a target image sequence including objects, determine position change information of each of the at least one object in the target image sequence; a crowd behavior detecting module 82, configured to, based on the position change information obtained in the target image sequence, perform graph convolution processing, and based on extracted features obtained by the graph convolution processing, determine a crowd behavior corresponding to the objects in the target image sequence.

[135] In some illustrated embodiments, the position change information determining module 81 is specifically configured to: perform image processing for each image in the target image sequence to determine position information of each of the at least one object in each image; perform object tracking for each of the at least one object so as to determine the position change information of each of the at least one object in the target image sequence based on a tracking result and the position information.

[136] In some illustrated embodiments, the position change information determining module 81 is specifically configured to: by using a Kalman filter algorithm or an object detection model, perform object tracking for each of the at least one object; based on the position information of a same tracked object in each image, determine the position change information of each of the at least one object.

[137] In some illustrated embodiments, the crowd behavior detecting module 82 includes: a spatial graph convolution processing module, configured to, based on position information of objects in each image in the target image sequence represented by the position change information and an association relationship between the objects in each image, perform spatial graph convolution processing for each image to obtain a graph feature corresponding to each image; a crowd behavior determining module, configured to perform time-domain convolution processing for the graph feature corresponding to each image, and based on extracted features obtained by the time-domain convolution processing, determine a crowd

behavior corresponding to the objects in the target image sequence; where the crowd behavior includes at least one of the followings: pedestrians gathering, pedestrians dispersion, pedestrians stranding, or pedestrians reversing.

[138] In some illustrated embodiments, the spatial graph convolution processing module is specifically configured to: based on the association relationship between objects in each image, determine a respective adjacency matrix corresponding to each image; based on the position information of the objects in each image, determine a respective feature matrix corresponding to each image; for each of image, based on the adjacency matrix corresponding to the image and the feature matrix corresponding to the image, performing the spatial graph convolution processing to obtain a graph feature corresponding to the image.

[139] In some illustrated embodiments, the apparatus 80 further includes an association relationship determining module, configured to determine an association relationship between any two objects included in each image in the target image sequence.

[140] In some illustrated embodiments, the association relationship determining module is specifically configured to: for each of the at least one object included in each image, extract an image feature corresponding to a region where the object is located in the image, where the image feature represents image information of a position where the object is located; based on the image feature corresponding to each object, determine a similarity between any two objects in the objects; determine two objects with a similarity greater than or equal to a first preset threshold as two objects having an association relationship.

[141] In some illustrated embodiments, the association relationship determining module is specifically configured to: respectively perform image processing for each image to determine position information of the objects in each image; based on the position information corresponding to each object, determine a distance between any two objects in the objects; based on the distance, determine the association relationship between the any two objects included in each image.

[142] In some illustrated embodiments, the association relationship determining module is specifically configured to: map the determined distance between any two objects to an interval formed by a third preset threshold and a fourth preset threshold; determine the mapped distance between the any two objects as an association weight between the any two objects; by using the association weight between the any two objects, indicate the association relationship between the any two objects.

[143] In some illustrated embodiments, the graph convolution processing is implemented by using a graph convolution classification model; where, an apparatus for training the graph convolution classification model includes: a generating module, configured to generate training samples, where each of training samples has position change information of objects and labeling information of a crowd behavior based on the position change information of the objects; a training module, configured to, based on the position change information and the labeling information of the crowd behavior, train a preset graph convolution model to obtain the graph convolution classification model.

[144] In some illustrated embodiments, when generating one of the training samples, the generating module is specifically configured to: based on a motion simulation platform, set motion modes corresponding to virtual objects; based on the motion modes, determine position change information corresponding to each of the virtual objects; determine a crowd behavior represented by the position change information corresponding to the virtual objects; based on the position change information and the crowd behavior represented by the position change information, generate the training sample.

[145] The embodiment of the apparatus for detecting a crowd behavior provided by the present disclosure may be applied to an electronic device. Correspondingly, the present disclosure provides an electronic device, including a processor and a memory configured to store instructions executable by the processor, where the processor is configured to invoke the executable instructions stored in the memory to implement the method of detecting a crowd behavior according to any one of the above embodiments.

[146] FIG. 9 is a schematic diagram illustrating a hardware structure of an electronic device according to the present disclosure.

[147] As shown in FIG. 9, the electronic device may include a processor for executing instructions, a network interface for performing network connection, a memory for storing running data for the processor, and a non-volatile memory storing instructions corresponding to the apparatus for detecting a crowd behavior.

[148] The apparatus embodiment may be implemented by software or hardware or a combination of software and hardware. With software implementation as an example, the apparatus, as a logical means, is formed by reading corresponding computer instructions in the non-volatile memory to the memory for running through a processor of an electronic device where the apparatus is located. From hardware level, in addition to the processor, the memory, the network interface, and the non-volatile memory shown in FIG. 9, the electronic

device where the apparatus is located usually may include other hardware based on actual functions of the electronic device in the embodiments, which is not redundantly described herein. It may be understood that in order to increase the processing speed, the instructions corresponding to the apparatus for detecting a crowd behavior may also be directly stored in the memory, which is not limited herein.

[149] The present disclosure provides a computer readable storage medium, which may be a volatile or non-volatile storage medium. The storage medium stores computer programs, where the computer programs are executed to implement the method of detecting a crowd behavior according to any one of the above embodiments.

[150] Those skilled in the art should understand that one or more embodiments of the present disclosure may be provided as methods, systems, or computer program products. Therefore, one or more embodiments of the present disclosure may take the form of a pure hardware implementation, a pure software implementation, or an implementation combining software and hardware. Furthermore, one or more embodiments of the present disclosure may take the form of a computer program product implemented on one or more computer available storage mediums (including but not limited to disk memories, CD-ROM, optical memories, etc.) containing computer available program codes.

[151] In the present disclosure, the word “and/or” refers to at least one of two. For example, “A and/or B” may include three options: A, B and “A and B”.

[152] Different embodiments in the present disclosure are described in a progressive manner. Each embodiment focuses on the differences from other embodiments with those same or similar parts among the embodiments referred to each other. Particularly, since data processing device embodiments are basically similar to the method embodiments, the device embodiments are briefly described with relevant parts referred to the descriptions of the method embodiments.

[153] Specific embodiments of the present disclosure are described above. Other embodiments not described herein still fall within the scope of the appended claims. In some cases, the actions or steps recorded in the claims may be performed in a sequence different from the embodiments to achieve a desired result. Further, the processes shown in drawings do not necessarily require a particular sequence or a continuous sequence to achieve the desired result. In some embodiments, a multi-task processing and parallel processing are possible and may also be advantageous.

[154] The embodiments of the subject and functional operations described in the present disclosure may be achieved in the following: a digital electronic circuit, a tangible computer software or firmware, a computer hardware including a structure disclosed in the present disclosure or a structural equivalent thereof, or a combination of one or more of the above. The embodiment of the subject described in the present disclosure may be implemented as one or more computer programs, that is, one or more modules in computer program instructions encoded on a tangible non-transitory program carrier for being executed by or controlling a data processing apparatus. Alternatively or additionally, the program instructions may be encoded on an artificially-generated transmission signal, such as a machine-generated electrical, optical or electromagnetic signal. The signal is generated to encode and transmit information to an appropriate receiver for execution by the data processing apparatus. The computer storage medium may be a machine readable storage device, a machine readable storage substrate, a random or serial access memory device, or a combination of one or more of the above.

[155] The processing and logic flows described in the present disclosure may be executed by one or more programmable computers executing one or more computer programs to perform operations based on input data and generate outputs to execute corresponding functions. The processing and logic flows may be further executed by a dedicated logic circuit, such as a field programmable gate array (FPGA) or an application specific integrated circuit (ASIC), and the apparatus may be further implemented as the dedicated logic circuit.

[156] Computers suitable for executing computer programs may include, for example, a general-purpose and/or special-purpose microprocessor, or any other type of central processing unit. Generally, the central processing unit receives instructions and data from a read-only memory and/or random access memory. Basic components of a computer may include a central processing unit for implementing or executing instructions and one or more storage devices for storing instructions and data. Generally, the computer may further include one or more mass storage devices for storing data, such as magnetic disks, magneto-optical disks or compact disks, or the computer is operably coupled to this mass storage device to receive data therefrom or transmit data thereto, or both. However, the computer does not necessarily have such device. In addition, the computer may be embedded in another device, such as a mobile phone, a personal digital assistant (PDA), a mobile audio or video player, a game console, a global positioning system (GPS) receiver, or a portable storage device, e.g., a universal serial bus (USB) flash drive, and so on.

[157] Computer readable media suitable for storing computer program instructions and data may include all forms of non-volatile memories, media and memory devices, such as semiconductor memory devices (e.g., Erasable Programmable Read-Only Memory (EPROM), Electrically Erasable Programmable Read-Only Memory (EEPROM) and flash memory device), magnetic disks (e.g., internal hard disk or removable disk), magneto-optical disks, and CD ROM and DVD-ROM disks. The processor and the memory may be supplemented by or incorporated into a dedicated logic circuit.

[158] Although many specific implementation details are included in the present disclosure, these details should not be construed as limiting any scope of the present disclosure or the claimed scope, but are mainly used to describe the features of specific embodiments of the present disclosure. Certain features described in several embodiments of the present disclosure may also be implemented in combination in a single embodiment. On the other hand, various features described in the single embodiment may also be implemented separately or in any appropriate sub-combination in several embodiments. In addition, although the features may function in certain combinations as described above and even be initially claimed as such, one or more features from the claimed combination may be removed from the combination in some cases, and the claimed combination may refer to a sub-combination or a variation of the sub-combination.

[159] Similarly, although the operations are described in a specific order in the drawings, this should not be understood as requiring these operations to be performed in the shown specific order or in sequence, or requiring all of the illustrated operations to be performed, so as to achieve a desired result. In some cases, multi-task processing and parallel processing may be advantageous. In addition, the separation of different system modules and components in the above embodiments should not be understood as requiring such separation in all embodiments. Further, it is to be understood that the described program components and systems may be generally integrated together in a single software product or packaged into a plurality of software products.

[160] Therefore, the specific embodiments of the subject are already described, and other embodiments are within the scope of the appended claims. In some cases, actions recorded in the claims may be performed in a different order to achieve the desired result. In addition, the processing described in the drawings is not necessarily performed in the shown specific order or in sequence, so as to achieve the desired result. In some implementations, multi-task processing and parallel processing may be advantageous.

[161] The foregoing disclosure is merely illustrative of one or more of preferred embodiments of the present disclosure but not intended to limit one or more embodiments of the present disclosure, and any modifications, equivalent substitutions, adaptations thereof made within the spirit and principles of one or more embodiments in the present disclosure shall be encompassed in the scope of protection of one or more embodiments in the present disclosure.

CLAIMS

1. A method of detecting a crowd behavior, comprising:

performing object tracking for at least one object in a target image sequence comprising objects to determine position change information of each of the at least one object in the target image sequence;

based on the position change information obtained in the target image sequence, performing graph convolution processing; and

based on extracted features obtained by the graph convolution processing, determining a crowd behavior corresponding to the objects in the target image sequence.

2. The method of claim 1, wherein performing object tracking for the at least one object in the target image sequence comprising the objects to determine the position change information of each of the at least one object in the target image sequence comprises:

performing image processing for each image in the target image sequence to determine position information of each of the at least one object in a corresponding image;

performing object tracking for each of the at least one object so as to determine the position change information of each of the at least one object in the target image sequence based on a tracking result and the position information.

3. The method of claim 2, wherein performing object tracking for each of the at least one object so as to determine the position change information of each of the at least one object in the target image sequence based on the tracking result and the position information comprises:

by using a Kalman filter algorithm or an object detection model, performing object tracking for each of the at least one object;

based on the position information of a same tracked object in corresponding images, determining the position change information of each of the at least one object.

4. The method of any one of claims 1 to 3, wherein based on the position change information obtained in the target image sequence, performing graph convolution processing to determine the crowd behavior corresponding to the objects in the target image sequence comprises:

based on position information of objects in at least one image in the target image sequence represented by the position change information and an association relationship between the objects in the at least one image, respectively performing spatial graph convolution processing for each of the at least one image to obtain a respective graph feature corresponding to each of the at least one image;

performing time-domain convolution processing for the graph feature corresponding to each of the at least one image, and based on extracted features obtained by the time-domain convolution processing, determining the crowd behavior corresponding to the objects in the target image sequence; wherein the crowd behavior comprises at least one of the followings: pedestrians gathering, pedestrians dispersion, pedestrians stranding, or pedestrians reversing.

5. The method of claim 4, wherein based on the position information of objects in the at least one image in the target image sequence represented by the position change information and the association relationship between the objects in the at least one image, performing the spatial graph convolution processing for each of the at least one image respectively to obtain the graph feature corresponding to each of the at least one image comprises:

based on the association relationship between the objects in the at least one image, determining a respective adjacency matrix corresponding to each of the at least one image;

based on the position information of the objects in the at least one image, determining a respective feature matrix corresponding to each of the at least one image;

for each of the at least one image, based on the adjacency matrix corresponding to the image and the feature matrix corresponding to the image, performing the spatial graph convolution processing to obtain a graph feature corresponding to the image.

6. The method of any one of claims 1 to 5, wherein before performing the step of performing graph convolution processing based on the position change information obtained in the target image sequence to obtain the extracted features corresponding to the target image sequence, the method further comprises:

determining an association relationship between any two objects comprised in at least one image in the target image sequence.

7. The method of claim 6, wherein determining the association relationship between any two objects comprised in at least one image in the target image sequence comprises:

for each of the at least one object comprised in the at least one image, extracting an image feature corresponding to a region where the object is located in the image, wherein the image feature represents image information of a position where the object is located;

based on the image feature corresponding to the at least one object, determining a similarity between any two objects in the at least one object;

determining two objects with a similarity less than a first preset threshold as two objects having an association relationship.

8. The method of claim 6, wherein determining the association relationship between any two objects comprised in at least one image in the target image sequence comprises:

respectively performing image processing for each of the at least one image to determine position information of the at least one object in the at least one image;

based on the position information corresponding to the at least one object, determining a distance between any two objects in the at least one object;

based on the distance, determining the association relationship between the any two objects comprised in the at least one image.

9. The method of claim 8, wherein based on the distance, determining the association relationship between the any two objects comprised in the at least one image comprises:

mapping the determined distance between the any two objects to an interval formed by a third preset threshold and a fourth preset threshold;

determining the mapped distance between the any two objects as an association weight between the any two objects;

by using the association weight between the any two objects, indicating the association relationship between the any two objects.

10. The method of any one of claims 1 to 9, wherein,

the graph convolution processing is implemented by using a graph convolution classification model;

wherein, the graph convolution classification model is trained by:

generating a training sample, wherein the training sample has position change information of objects and labeling information of a crowd behavior based on the position change information of the objects;

based on the position change information and the labeling information of the crowd behavior, training a preset graph convolution model to obtain the graph convolution classification model.

11. The method of claim 10, wherein generating the training sample comprises:

based on a motion simulation platform, setting motion modes corresponding to virtual objects;

based on the motion modes, determining position change information corresponding to at least one of the virtual objects;

determining a crowd behavior represented by the position change information corresponding to the at least one of the virtual objects;

based on the position change information and the crowd behavior represented by the position change information, generating the training sample.

12. An apparatus for detecting a crowd behavior, comprising:

a position change information determining module, configured to, based on an object tracking result of at least one object in a target image sequence comprising objects, determine position change information of each of the at least one object in the target image sequence;

a crowd behavior detecting module, configured to, based on the position change information obtained in the target image sequence, perform graph convolution processing; and based on extracted features obtained by the graph convolution processing, determine a crowd behavior corresponding to the objects in the target image sequence.

13. An electronic device, comprising:

a processor;

a memory storing instructions executable by the processor;

wherein, the processor is configured to invoke the executable instructions stored in the memory to implement the method of detecting a crowd behavior according to any one of claims 1 to 11.

14. A computer readable storage medium, storing computer programs, wherein the computer programs are executed to implement the method of detecting a crowd behavior according to any one of claims 1 to 11.

15. A computer program product, wherein the computer program product is run on a computer to cause the computer to implement the method of detecting a crowd behavior according to any one of claims 1 to 11.