(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 November 2024 (28.11.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/239207 A1

(51) International Patent Classification:
G06T 3/00 (2006.01) *G06T 7/00* (2017.01)

(21) International Application Number:
PCT/CN2023/095578

(22) International Filing Date:
22 May 2023 (22.05.2023)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for CN only)*: **NOKIA SHANGHAI BELL CO., LTD.** [CN/CN]; 388 Ning Qiao Road, China (Shanghai) Pilot Free Trade Zone, Pudong New Area, Shanghai 201206 (CN).

(71) Applicant: **NOKIA SOLUTIONS AND NETWORKS OY** [FI/FI]; Karakaari 7, Espoo, 02610 (FI).

(72) Inventors: **LI, Liu Hai**; 388 Ning Qiao Rd, Pudong New Area, Shanghai 201206 (CN). **SUN, Yu Feng**; 388 Ning Qiao Rd, Pudong New Area, Shanghai 201206 (CN). **CHEN, Wei**; 388 Ning Qiao Rd, Pudong New Area, Shanghai 201206 (CN). **YUAN, Yong Fu**; 388 Ning Qiao Rd, Pudong New Area, Shanghai 201206 (CN).

(74) Agent: **BEIJING SHIHUI LAW FIRM**; 42/F, Tower C, Beijing Yintai Centre, No. 2 Jianguomenwai Avenue, Chaoyang District, Beijing 100022 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA,

(54) Title: THREE-DIMENSIONAL OBJECT DISTANCE ESTIMATION

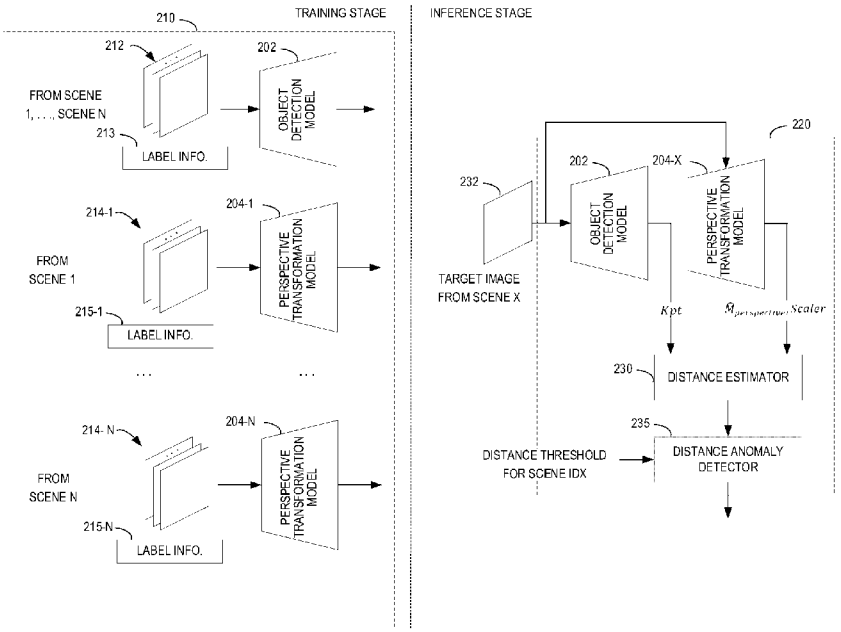


FIG. 2

(57) **Abstract:** Example embodiments of the present disclosure relate to three-dimensional (3D) object distance estimation. According to a method, a first location of a first object and a second location of a second object are detected within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene. A perspective transformation and a distance scale are determined for the target image by providing the target image into a perspective transformation model associated with the target deployment. Based on the determined perspective transformation, the first and the second locations are transformed into a first transformed location and a second transformed location within a top plane view of a three-dimensional space, respectively. A target distance between the first and the second objects in a three-dimensional space is determined by scaling a distance between the first and the second transformed location with the distance scale.

WO 2024/239207 A1

[Continued on next page]

NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

THREE-DIMENSIONAL OBJECT DISTANCE ESTIMATION

FIELDS

[0001] Various example embodiments of the present disclosure generally relate to the field of telecommunication and in particular, to methods, devices, apparatuses and computer readable storage medium for three-dimensional (3D) object distance estimation.

BACKGROUND

[0002] In various applications, there is a need to estimate 3D distances between objects in an environment. For example, to avoid a hazard event (e.g., a collision) between two objects, it needs to estimate a distance between the two objects and determine whether the distance is too low (lower than a threshold). At present, artificial intelligence technology and imaging control technology are developing rapidly, and a large number of models have been implemented with the model architecture based on various neural networks, such as target detection models, image generation models, and the like. To estimate 3D object distances, the main challenge is how to accurately restore the 3D scene.

SUMMARY

[0003] In a first aspect of the present disclosure, there is provided an apparatus. The apparatus comprises at least one processor; and at least one memory storing instructions that, when executed by the at least one processor, cause the apparatus at least to: detect a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene; determine a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment; transform, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and determine a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

[0004] In a second aspect of the present disclosure, there is provided a method. The method comprises: detecting a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene; determining a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment; transforming, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and determining a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

[0005] In a third aspect of the present disclosure, there is provided an apparatus. The apparatus comprises means for detecting a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene; means for determining a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment; means for transforming, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and means for determining a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

[0006] In a fourth aspect of the present disclosure, there is provided a computer readable medium. The computer readable medium comprises instructions stored thereon for causing an apparatus to perform at least the method according to the second aspect.

[0007] It is to be understood that the Summary section is not intended to identify key or essential features of embodiments of the present disclosure, nor is it intended to be used to limit the scope of the present disclosure. Other features of the present disclosure will become easily comprehensible through the following description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] Some example embodiments will now be described with reference to the accompanying drawings, where:

[0009] FIG. 1 illustrates an example environment in which example embodiments of the present disclosure can be implemented;

[0010] FIG. 2 illustrates a schematic diagram of a framework for model-based object distance estimation in accordance with some example embodiments of the present disclosure;

[0011] FIG. 3 illustrates a flowchart of a process for model training and inference for model-based object distance estimation in accordance with some example embodiments of the present disclosure;

[0012] FIG. 4 illustrates a flowchart of a process for training the object detection model in the framework of FIG. 2 in accordance with some example embodiments of the present disclosure;

[0013] FIG. 5 illustrates a flowchart of a process for training the perspective transformation model in the framework of FIG. 2 in accordance with some example embodiments of the present disclosure;

[0014] FIG. 6A and FIG. 6B illustrate schematic diagrams of example image transformation and normalized object distances in a 3D top plane view in accordance with some example embodiments of the present disclosure;

[0015] FIG. 7 illustrates a processing flow for training the perspective transformation model in the framework of FIG. 2 in accordance with some example embodiments of the present disclosure;

[0016] FIG. 8 illustrates a schematic diagram of floating key points in the model training in accordance with some example embodiments of the present disclosure;

[0017] FIG. 9 illustrates a flowchart of a process for object distance estimation and abnormal distance detection in an inference stage in accordance with some example embodiments of the present disclosure;

[0018] FIG. 10 illustrates a simplified block diagram of a device that is suitable for implementing example embodiments of the present disclosure; and

[0019] FIG. 11 illustrates a block diagram of an example computer readable medium in accordance with some example embodiments of the present disclosure.

[0020] Throughout the drawings, the same or similar reference numerals represent the same or similar element.

DETAILED DESCRIPTION

[0021] Principle of the present disclosure will now be described with reference to some example embodiments. It is to be understood that these embodiments are described only for the purpose of illustration and help those skilled in the art to understand and implement the present disclosure, without suggesting any limitation as to the scope of the disclosure. Embodiments described herein can be implemented in various manners other than the ones described below.

[0022] In the following description and claims, unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skills in the art to which this disclosure belongs.

[0023] References in the present disclosure to "one embodiment," "an embodiment," "an example embodiment," and the like indicate that the embodiment described may include a particular feature, structure, or characteristic, but it is not necessary that every embodiment includes the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0024] It shall be understood that although the terms "first," "second" and the like may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first element could be termed a second element, and similarly, a second element could be termed a first element, without departing from the scope of example embodiments. As used herein, the term "and/or" includes any and all combinations of one or more of the listed terms.

[0025] As used herein, "at least one of the following: <a list of two or more elements>"

and "at least one of <a list of two or more elements>" and similar wording, where the list of two or more elements are joined by "and" or "or", mean at least any one of the elements, or at least any two or more of the elements, or at least all the elements.

[0026] As used herein, unless stated explicitly, performing a step "in response to A" does not indicate that the step is performed immediately after "A" occurs and one or more intervening steps may be included.

[0027] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of example embodiments. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises", "comprising", "has", "having", "includes" and/or "including", when used herein, specify the presence of stated features, elements, and/or components etc., but do not preclude the presence or addition of one or more other features, elements, components and/ or combinations thereof.

[0028] As used in this application, the term "circuitry" may refer to one or more or all of the following:

(a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry) and

(b) combinations of hardware circuits and software, such as (as applicable):

(i) a combination of analog and/or digital hardware circuit(s) with software/firmware and

(ii) any portions of hardware processor(s) with software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile phone or server, to perform various functions) and

(c) hardware circuit(s) and or processor(s), such as a microprocessor(s) or a portion of a microprocessor(s), that requires software (e.g., firmware) for operation, but the software may not be present when it is not needed for operation.

[0029] This definition of circuitry applies to all uses of this term in this application,

including in any claims. As a further example, as used in this application, the term circuitry also covers an implementation of merely a hardware circuit or processor (or multiple processors) or portion of a hardware circuit or processor and its (or their) accompanying software and/or firmware. The term circuitry also covers, for example and if applicable to the particular claim element, a baseband integrated circuit or processor integrated circuit for a mobile device or a similar integrated circuit in server, a cellular network device, or other computing or network device.

[0030] Communication in the present disclosure can follow any suitable type of communication standards, including cellular communication, Wi-Fi communication, LPWAN communication, data communication, voice communication, multimedia communication, short-range communications such as Bluetooth, near-field communication, location-based communication such as the use of the global positioning system (GPS) to determine a location, another like communication function, or any combination thereof. Communications may be implemented in according to one or more communication protocols and/or standards, such as IEEE 802.11, Code Division Multiplexing Access (CDMA), Wideband Code Division Multiple Access (WCDMA), GSM, LTE, New Radio (NR), UMTS, WiMax, Ethernet, transmission control protocol/internet protocol (TCP/IP), synchronous optical networking (SONET), Asynchronous Transfer Mode (ATM), QUIC, Hypertext Transfer Protocol (HTTP), and so forth.

[0031] As used herein, the term "model" is referred to as an association between an input and an output learned from training data, and thus a corresponding output may be generated for a given input after the training. The generation of the model may be based on a machine learning technique. The machine learning techniques may also be referred to as artificial intelligence (AI) techniques. In general, a machine learning model can be built, which receives input information and makes predictions based on the input information. For example, a classification model may predict a class of the input information among a predetermined set of classes. As used herein, "model" may also be referred to as "machine learning model", "learning model", "machine learning network" , or "learning network," which are used interchangeably herein.

[0032] Generally, machine learning may usually involve three stages, i.e., a training stage, a validation stage, and an inference stage (also referred to as an inference stage).

At the training stage, a given machine learning model may be trained (or optimized) iteratively using a great amount of training data until the model can obtain, from the training data, consistent inference similar to those that human intelligence can make. During the training, a set of parameter values of the model is iteratively updated until a training objective is reached. Through the training process, the machine learning model may be regarded as being capable of learning the association between the input and the output (also referred to an input-output mapping) from the training data. At the validation stage, a validation input is applied to the trained machine learning model to test whether the model can provide a correct output, so as to determine the performance of the model. Generally, the validation stage may be considered as a step in a training process, or sometimes may be omitted. At the inference stage, the resulting machine learning model may be used to process a real-world model input based on the set of parameter values obtained from the training process and to determine the corresponding model output.

[0033] For 3D distance estimation, there is a disagreement in whether the scene can be accurately restored to a 3D space by single-view two-dimensional (2D) images. In this context, a majority of works aim at accurate 3D scene reduction by adding more information, including, for example, Light Detection And Ranging () sensing data, images captured by multi-perspective camera, and the like. For example, some solutions propose to obtain the multimodal data of 3D scene through the LIDAR-based sensor devices or imaging devices, combine the multimodal data to perform data fusion, and then generate a complete 3D scene for object distance estimation. Some other solution propose to directly estimate uncertain 3D objects by two-dimensional image data through a trained model, where the estimate is made by combining multi-perspective cameras with conversion of the 3D scene. However, those solutions often significantly increase the design and inference cost of the model as well as high costs of hardware device for data collection.

[0034] In the case where images are collected by monocular cameras, the following data are missing for 2D-to-3D object scene conversion: depth detection data in the images, and the intrinsic and extrinsic information corresponding to the camera. In order to collect the above data, the cost is large or even impossible in the real application scenes. Therefore, it is desired to effectively estimate 3D object distance through low-cost manners.

[0035] According to example embodiments of the present disclosure, there is proposed a solution for 3D object distance estimation from 2D images. This solution is based on

ML models. Specifically, a perspective transformation model is trained for a deployment scene of a camera and is used to determine a perspective transformation and a distance scale for a specific image captured by the camera in the deployment scene. For a first object and a second object detected in a target image, the perspective transformation is used to transform a first location of a first object and a second location of a second object within a 2D space corresponding to a target image into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a 3D space, respectively. A target distance between the first object and the second object in a three-dimensional space is then determined by scaling a distance between the first transformed location and the second transformed location with the distance scale.

[0036] The present solution can effectively utilize the 2D image and greatly reduce the modeling and usage costs for achieving 2D-to-3D scene restoration by increasing the information perception dimension (multi-vision camera, LIDAR, etc.) in the application field. Further, the present solution does not require distance metric calculation on the basis of 3D model, and ca estimates the plane distance of 2D object in the 3D space directly by model reduction, which greatly reduces the technical development threshold and usage cost in the application field, and enhances the convenience in different application scenarios. As such, the present solution can be easily adopted in light application scenarios.

[0037] Example embodiments of the present disclosure will be described in detail below with reference to the accompanying drawings. It is noted that any section/subsection headings provided herein are not intended to be limiting. Embodiments are described throughout this document, and any type of embodiment may be included under any section/subsection. Furthermore, embodiments disclosed in any section/subsection may be combined with any other embodiments described in the same section/subsection and/or a different section/subsection in any manner.

Example Environment

[0038] FIG. 1 illustrates an example environment 100 in which example embodiments of the present disclosure can be implemented. In the environment 100, a camera is mounted in a physical environment to capture images. As shown, a camera 102-1 is mounted with a specific deployment in a scene 101-1 (represented as "Scene 1"), ..., a camera 102-N is mounted with a specific deployment in a scene 101-N (represented as "Scene N"), where N may be an integer larger than or equal to one. The cameras 102-1, ...,

102-N may be the same or different types of cameras, e.g., with different imaging parameters. The deployments of the cameras 102-1, ..., 102-N may be the same or different although the scenes may be different. In some example embodiments of the present disclosures, the cameras 102-1, ..., 102-N may be monocular images, capturing monochrome images or colored images (such as RGB images).

[0039] For the purpose of discussion, the cameras 102-1, ..., 102-N may be collectively or individually referred to as cameras 102, and the scenes 101-1, ..., 101-N may be e collectively or individually referred to as scenes 101.

[0040] In embodiments of the present disclosure, the images captured by the cameras 102 may be transmitted to a device(s)/system(s) for object distance estimation and in some cases, for abnormal distance detection. The images may be transmitted via various networking technology. For example, the camera 102-1 may connect to a radio access network (RAN) and thus can transmit its images to a network device 103 in the RAN. The network device 103 may convey the images to a core transmission network 110, e.g., via a switch 104 and an access router 105. The images may be routed to an object distance estimation device in an information management monitoring room 120 which can access to the core transmission network 110 via an access router 122. The camera 102-N may have a network connection in a local access network (LAN) and thus can transmit its images to the core transmission network 110 via a switch 106 in the LAN. The images captured by the camera 102-N may also be conveyed to an object distance estimation device in the information management monitoring room 120.

[0041] In some example embodiments, the devices in the information management monitoring room 120 may subscribe to remote computing resources from a computing center 130 for processing the images and/or other processing operations required in the object distance estimation and the abnormal distance detection. The computing center 130 may be accessed through the core transmission network 110 via an access router 132. The computing center 130 may contain one or more computing resource clusters to provide the processing, memory, storage, and networking resources for the information management monitoring room 120. In some example embodiments, the computing center 130 may be implemented as a cloud platform.

[0042] It would be appreciated that in some example embodiments, images captured by a camera 102 may be processed remotely (e.g., in the information management monitoring

room 120 and/or in the computing center 130) for the object distance estimation and abnormal distance detection, and in some other embodiments, the images captured by the camera 102 may be processed locally, e.g., by a device deployed locally in the scene and communicatively coupled with the camera 102. The scope of the present disclosure is not limited in this regard.

Work Principle and Example Framework

[0043] The example embodiments of the present disclosure propose a solution to derive 3D distance information from an 2D image (e.g., an image captured by a monocular camera) using a perspective transformation model. The perspective transformation model is trained for a target deployment of a target camera, and is used to generate a perspective transformation and a distance scale for a target image captured by the target camera. In some example embodiments, as will described in detail below, an object detection model may be trained to detect the objects from the target image. As the models are introduced, there may be a training stage for the models and then an inference stage to apply the trained models for the distance estimation.

[0044] FIG. 2 illustrates a framework 200 for model-based object distance estimation in accordance with some example embodiments of the present disclosure. The framework 200 comprises a model training system 210 in a training stage and a model inference system 220 in an inference stage. The model training system 210 is configured to train one or more perspective transformation models 204-1, ..., 204-N for one or more scenes (Scene 1, ..., Scene N) respectively. For the purpose of discussion, the perspective transformation models 204-1, ..., 204-N may be collectively or individually referred to as perspective transformation models 204. In some example embodiments, the model training system 210 may be further configured to train an object detection model 202 to detect an object(s) in an image. The model inference system 220 is configured to determine a distance between objects detected from a target image 232 using the trained object detection model 202 and the trained perspective transformation model 204.

[0045] The model training system 210 and the model inference system 220 may be the same or different systems. For example, in the example environment 100 as shown in FIG. 1, the model training system 210 may be implemented at the computing center 130, and the model inference system 220 may be implemented at one or more devices in the information management monitoring room 120. In another example, both the model

training system 210 and the model inference system 220 may be implemented as a same system/device in the computing center 130 and/or in the information management monitoring room 120.

[0046] The object detection model 202 and the perspective transformation models 204 are first trained and then applied for inference. FIG. 3 illustrates a flowchart of a process 300 for model training and inference for model-based object distance estimation in accordance with some example embodiments of the present disclosure. The process 300 will be described with reference to FIG. 2. Some training-related operations in the process 300 may be performed by the model training system 210, and some inference-related operations in the process 300 may be performed by the model inference system 220.

[0047] At block 310, the model training system 210 collects first sample images 212 and first label information 213 indicating objects detected in the first sample images 212. At block 320, the model training system 210 trains the object detection model 202 with the first sample images 212 and first label information 213.

[0048] The input to the object detection model 202 may be an image, and the output from the object detection model 202 may be a location(s) of an object(s) in the input image. In some example embodiments, the object detection model 202 may provide a bounding box to locate an object in the image.

[0049] As shown in FIG. 3, the training data for the object detection model 202 may comprise sample images 212 (sometimes referred to as first sample images 212) and first label information 213 indicating objects detected from the first sample images 212. The object detection model 202 is used to detect one or more objects of interest from an image (the distance(s) of which are to be estimated). Thus it is expected that the object detection 202 may not be associated with a specific scene or a deployment of a camera in the scene. In some example embodiments, the first sample images 212 may be collected from various sources, for example, by a plurality of cameras with a plurality of deployments in Scene 1, ..., Scene N.

[0050] The training of the object detection model 202 will be discussed in more detail below with reference to FIG. 4.

[0051] At block 330, the model training system 210 collects second sample images 214-1, ..., 214-N and second label information 215-1, ..., 215-N for the second sample images

214-1, ..., 214-N. At block 340, the model training system 210 trains the perspective transformation model 204-1, ..., 204-N with the second sample images 214-1, ..., 214-N and second label information 215-1, ..., 215-N, respectively.

[0052] For the purpose of discussion, the second sample images 214-1, ..., 214-N may be collectively or individually referred to as second sample images 214; and the second label information 215-1, ..., 215-N may be collectively or individually referred to as second label information 215.

[0053] The input to a perspective transformation model 204 may be an image, with one or more objects detected in the image, and the output from the perspective transformation model 204 may be a perspective transformation and a distance scale for the input image. The perspective transformation is used to transform the input image (which is in the 2D pixel space) into a top plane view of a 3D space, to obtain a transformed image. The distance scale is used to scale distances between pixels (or object locations) in the transformed image into the physical distance in the scene. The perspective transformation model 204 may also be referred to as a "perspective transformation and filed of view size" model.

[0054] Each perspective transformation model 204 may be associated with a scene, and it is assumed that a camera is deployed in the scene with a specific deployment. The perspective transformation model 204 is trained to adopt to the perspective transformation and field of view scaling for the specific deployment of the camera, and thus can be used for inference of images captured by a camera with the corresponding deployment. Specifically, training data for a perspective transformation model 204-1 may include second sample images 214-1 captured by one or more cameras with a deployment (Deployment 1) in Scene 1, ..., training data for a perspective transformation model 204-N may include second sample images 214-N captured by one or more cameras with a different deployment (Deployment N) in Scene N. In some example embodiments, the second sample images 214-1, ..., 214-N used for training the perspective transformation models may be divided from the first sample images 212 according to the scenes or deployments.

[0055] In some example embodiments, a deployment of a camera may be identified by a mounting height, a shooting orientation, and a type of the camera in the scene. For example, for the same type of image, if it is deployed at different heights and/or with

different orientations, then the modelling for 2D-to-3D scene restoration may be different. For different types of cameras, it may have different focal lengths, and then the modelling for 2D-to-3D scene restoration may be different. Then different perspective transformation models 204 may need to be trained for the different deployments of cameras.

[0056] The training data for a perspective transformation model 204 may further include second label information 215 for the second sample images 214. For each second sample image 214, the second label information 215 may indicate a labeled indication of whether a distance between a pair of objects detected in the second sample image is abnormal. In some example embodiments, for a second sample image 214, the second label information may further indicate labeled transformed locations of the pair of objects in the top plane view.

[0057] The training of the perspective transformation models 204 will be described in detail below with reference to FIG. 5.

[0058] With the object detection model 202 and the perspective transformation models 204 have been trained, at block 350, the model inference system 220 determines a target distance between a first object and a second object in a target image 232 using the object detection model 202 and a perspective transformation model 204. As shown in FIG.2, the model inference system 220 may comprise a distance estimator 230 to determine the target distance between the first and second objects using the outputs from the object detection model 202 and the perspective transformation model 204.

[0059] The perspective transformation model 204 applied for the target image 232 may depend on a target deployment of a target camera which captures the target image 232. It is assumed that the target image 232 is captured by a target camera with a target deployment in Scene X. Among all the perspective transformation models 204, the perspective transformation model 204-X associated with the target deployment in Scene X may be apply by the model inference system 220 for determining the distance.

[0060] The object detection model 202 is used to detect two or more objects from the target image 232, to determine locations of the two or more objects in the target image 232. The locations may represent location information of the objects within a 2D space (also referred to as a 2D pixel space) corresponding to the target image 232. The perspective transformation model 204-X is used to transform the location of a first object

and the location of a second object within the 2D space into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a 2D space, respectively. The transformed locations are considered as 3D locations for the first and second objects.

[0061] In some example embodiments, at block 360, the model inference system 220 may perform abnormal distance detection based on the target distance between the first object and the second object. As shown in FIG.2, the model inference system 220 may comprise an distance anomaly detector 235 to perform the abnormal distance detection. The distance anomaly detector 235 may determine an indication of whether the target distance between the first object and the second object is abnormal based on a comparison result between the target distance and a distance threshold.

[0062] The object distance estimation and abnormal distance detection will be further discussed in more detail below with reference to FIG. 9.

[0063] The present solution of the example embodiments disclosure herein can effectively improve the efficiency of 2D imaging to 3D distance detection, greatly reduce the hardware cost and model training and inference costs for scenarios such as in-site hazard distance detection, while achieving the expected 3D distance detection in multiple crossover application areas, including but not limited to pedestrian distance detection, construction vehicle safety distance detection, hazardous object safety distance detection, and the like.

Training of Object Detection Model

[0064] FIG. 4 illustrates a flowchart of a process 400 for training the object detection model 202 in the framework 200 of FIG. 2 in accordance with some example embodiments of the present disclosure. For the purpose of discussion, the process 400 will be described from the perspective of the model training system 210 in FIG. 2.

[0065] At block 410, the model training system 210 obtains first sample images captured by a plurality of cameras with a plurality of deployments and first label information indicating objects in the first sample images, for example, the first sample images 212 and the first label information 213 as shown in FIG. 2.

[0066] In some example embodiments, in order to train a generalized object detection model, the first sample images 212 may be collected from different dataset sources (for

example, captured by different cameras with different deployments in the scenes). In some example embodiments, the cameras used for capturing the first sample images 212 may be monocular cameras. In some example embodiments, a first sample image 212 may be stored as (scene ID, img_x), where the scene ID is used to identify a deployment of a camera in the scene, and one scene or deployment may correspond to one camera.

[0067] To ensure the generality of the first sample images 212 for training the object detection model 202, the differential data collection for multiple environmental states of the scenes may be considered, including but not limited to, different weather conditions (e.g., sunny, cloudy, rainy, etc.), different time conditions (e.g., early morning, midday, evening, etc.), and the like. If a specific deployment of a camera or the corresponding scene is a special deployment or a special scene, the impact of the environmental changes in that scene on the collection of the sample images may need to be taken into consideration.

[0068] The first sample images 212 are used to perform the training of the object detection model 202 which is primarily configured to localize objects in 2D pixel space. The first sample images 212 may be labeled with the objects occurred therein. In some example embodiments, the first sample images 212 may be labeled with bounding boxes of the objects to be detected, respectively. The labelling may be performed through manual annotations or through any suitable automation tools, to generate the first label information 213 for the first sample images 212. For an object labeled in a first sample images 212, the first label information 213 may indicate a bounding box of the area of the object in the first sample images 212. In some examples, the first label information 213 may include coordinates of the detected object in the first sample images 212, such as the four vertexes of the bounding box,, including the left, top, right, bottom vertexes (in this context, denotes l as left, t as top, r as right, b as bottom) for the pixel coordinate points.

[0069] In some example embodiments, for some use cases, it is expected to estimate a distance(s) between some specific types of objects, e.g., to evaluate whether those objects are too close to each other or too far away from each other. In those cases, the first sample images 212 containing the specific types of objects of interest may be sampled from all the available sample images. For example, if it is expected to determine an abnormal distance between a forklift and a pedestrian in a scene monitored by a camera, then the first sample images 212 containing a forklift and one or more pedestrians may be sampled from all the images captured by the camera. It would be appreciated that the object

detection model 202 may be trained to detect various types of objects at least including some types of objects whose distances are to be estimated, and in this case, more objects may be labeled in the first sample images 212.

[0070] At block 420, the model training system 210 constructs the object detection model 202. In some example embodiments, the object detection model 202 may be constructed using a generic detection model framework in computer vision. The scope of the present disclosure is not limited in the specific model structure for the object detection model.

[0071] At block 430, the model training system 210 trains the object detection model 202 with the first sample images 212 and the first label information 213.

[0072] During the training, the model training system 210 may iteratively optimize the object detection model 202 for completing the object detection task. The model training system 210 may apply any suitable model training algorithms and strategies to complete the training of the object detection model 202.

[0073] In some example embodiments, some hyperparameters may be set for the training of the object detection model 202. For example, a hyperparameter confidence_threshold may be set to a predetermined value (in a range between 0 and 1) for thresholding the selection of detected pre-selected bounding boxes of objects. Another hyperparameter NMS_IOU_threshold may be typically set to a predetermined value between 0 and 1 for controlling the screening of the pre-selected bounding boxes with high overlapping.

[0074] In some example embodiments, the iterative optimization may be completed if the F1-score of the object detection model 202 meets a training objective, e.g., the F1-score being close to 1 on a training dataset and the generalization of the object detection model 202 can prevent from overfitting on a validate dataset.

[0075] In some example embodiments, the first sample images 212 used for training the object detection model 202 may be further inspected, to guarantee the accuracy of the next model construction, and removing the missed and over detected RGB images from the sample data. After the training of the object detection model 202 is completed, the first sample images 212 may be evaluated to determine which objects are missed or mis-checked by the trained object detection model 202. If it is obvious that the object detection

model 202 has not yet trained to converge, re-training of the object detection model 202 may be performed with the first sample images 212 whose objects are missed. During the re-training, the gradient weights for those first sample images 212 may be enhanced. The model training may be iterated and the object detection model whose F1-score is at or near 1 on the training dataset and which has a strong generalization in the validation dataset, may be selected as the trained model for inference.

Training of Perspective Transformation Model

[0076] FIG. 5 illustrates a flowchart of a process 500 for training the perspective transformation model(s) 204 in the framework 200 of FIG. 2 in accordance with some example embodiments of the present disclosure. For the purpose of discussion, the process 600 will be described from the perspective of the model training system 210 in FIG. 2.

[0077] As mentioned above, the training data for a perspective transformation model 204 associated with a specific deployment of a scene may include second sample images 214 and second label information 215 for the second sample images 214. The training of one perspective transformation model 204 is discussed below and all the other perspective transformation models for other scenes may be trained in similar ways.

[0078] At block 510, the model training system 210 obtains the second sample images 214 with objects detected.

[0079] In some example embodiments, the second sample images 214 used for training a perspective transformation model 204 for a specific scene may include a subset of the first sample images 212 used for training the object detection model 202 which is collected from the specific scene. The first image samples 212 may be grouped according to the scene IDs, and the second sample images 214 may be selected from each scene (or deployment of the camera).

[0080] In some example embodiments, a part or all of the second sample images 214 may be different from those sample images used for training the object detection model 202. Then the objects in the second sample images 214 may be separately labeled from the labeling of the first sample images 212. In some example embodiments, if the object detection model 202 has been trained, the objects in the second sample images 214 may be detected using the trained object detection model 202.

[0081] In some example embodiments, the objects labeled in the second sample images

214 may be the types of objects whose distance(s) are to be estimated or the abnormal distance(s) are to be detected in the corresponding scene. For example, in one scene, it is expected to estimate a distance between a forklift and a pedestrian, and then these two types of objects are labeled in the second sample images 214 captured from this scene. For another scene, it is expected to estimate a distance between two cars, and then a type of car object is labeled in the second sample images 214 captured from the other scene.

[0082] At block 520, the model training system 210 determines a fitted perspective transformation based on the second sample images 214 with the objects detected. At block 530, the model training system 210 determines second label information 215 for the second sample images 214 based at least on the fitted perspective transformation. For each second sample image 214, the second label information 215 may indicate a labeled indication of whether a distance between a pair of objects detected in the second sample image is abnormal. In some example embodiments, for a second sample image 214, the second label information may further indicate labeled transformed locations of the pair of objects in the top plane view.

[0083] For the anomaly indication, in some example embodiments, in conjunction with labeling of the bounding boxes of the objects, the distances between one or more pairs of objects are also labeled for anomalies. For example, the annotators may evaluate whether the distance between a pair of objects in a second sample image 214 is abnormal, e.g., according to some reference objects or reference distance in the second sample image 214. The second sample image 214 and its second label information 215 may be stored as (scene ID, img_x, an indication of whether the distance is abnormal or not).

[0084] For the labeled transformed locations, a perspective transformation may be fitted from the labeled locations of the objects in the second sample images 214, for training of the perspective transformation model 204. As mentioned above, the perspective transformation model 204 is configured to output a perspective transformation and a distance scale for a specific deployment in a scene. The perspective transformation is used to transform a 2D location of an object in the 2D pixel space into a 3D location in a 3D space, or more specifically, a planar location in a top plane view of the 3D space. The distance scale is used to scale a distance between transformed location in the top plane view into the actual physical distance in the 3D space.

[0085] A perspective transformation (also referred to as a perspective transformation matrix)

may be represented as $M_{perspective} \in \mathbb{R}^{3\times3}$, the distance scale may be represented as $Scaler$. The $M_{perspective}$ may be flattened, for example, as a (1, 9) matrix, and the distance scale $Scaler$ is a (1, 1) matrix. A position transformation for a location in the 2D space corresponding to an image (e.g., a 2D pixel coordinate point) is defined as follows:

$$pts_{co-bev} = pts_{co-pix} \times M_{perspective} \quad (1)$$

where pts_{co-pix} represents the location in the image (e.g., the pixel coordinates in the image), and pts_{co-bev} represents a transformed location (e.g., the 2D top plane coordinates) in the 3D top plane view. The transformation may allow the 2D top plane coordinates obeying three degrees of freedom. Further, in order to keep the scale consistency, the distance scaler $Scaler$ is introduced, and then the physical distance between two objects (object *i* and *j*) in an image may be calculated as follows:

$$distance = \left\| pts_{co-bev}^{(i)} - pts_{co-bev}^{(j)} \right\|_{l2} \cdot Scaler \quad (2)$$

where $pts_{co-bev}^{(i)}$ represents a transformed location of object i in the top plane view, $pts_{co-bev}^{(j)}$ represents a transformed location of object *j* in the top plane view, $\| \quad \|_{l2}$ represents the L2-distance calculated from the 2D top plane coordinates of the two objects in the tope plane view.

[0086] FIG. 6A and FIG. 6B illustrate schematic diagrams of example image transformation and normalized object distances in a 3D top plane view. As shown, a captured image 610 or 640, which may be a 2D image captured by a monocular camera, may be transformed through a corresponding perspective transformation into a transformed image 620 or 650 which is considered as a view of the 3D scene from the tope plane view. The locations of objects 611, 612, and 613 in the image 610 may be mapped to respective transformed locations 621, 622, and 623 of the corresponding objects within the transformed image 620. After scale normalization, the transformed locations 621, 622, and 623 of the objects may be represented in a normalized top plane view 630. Similarly, the locations of objects 641, 642, and 643 in the image 640 may be mapped to respective transformed locations 651, 652, and 653 of the corresponding objects within the transformed image 650. After normalization, the transformed locations 651, 652, and 653 of the objects may be represented in a normalized top plane view 660.

[0087] The training of the perspective transformation model 204 is to learn an accurate perspective transformation as well as an accurate distance scale to estimate the 3D distance between two objects. For this purpose, in the training stage, a fitted perspective

transformation may be determined for a specific deployment in a scene. In some example embodiments, the locations of the objects labeled in the second sample images 214 (e.g., coordinates of four vertexes of a bounding box of an object in a second sample image 214) may be selected and used to calculate the fitted perspective transformation of the scene.

[0088] The fitted perspective transformation is used for transforming the locations of the objects in the second sample images 214 into transformed locations in the top plane view of the 3D space. The fitting process may be considered as there are errors in the manual evaluation, so a multi-labeling cross-validation approach is considered. Take a scene *i* as an example. One of the second sample images 214 collected for the scene *i* may be randomly selected for construction of the fitted perspective transformation. Another second sample image 214 is randomly selected for verifying whether the perspective transformation meets the expectations of the transformation. More specifically, for a selected second sample image 214, coordinates of four foot points in this image may be recorded as $[pts_1, pts_2, pts_3, pts_4]$, and the coordinates of the 4 foot points after the perspective transformation are defined as $[pts_1', pts_2', pts_3', pts_4']$. The transformed location(s) of one or more object(s) in the second sample image 214 (e.g., the coordinates of the four vertexes of the bounding box) may be evaluated to determine whether the locations (e.g., the coordinates) match the expected coordinates in the top view plane. The coordinates of the four foot points after the perspective transformation may be adjusted until they meet the expectations of the perspective transformation. That is, the fitting is repeated until the obtained fitted perspective transformation meets the expectation.

[0089] The fitted perspective transformation of each scene (Scene n) is recorded as (scene ID, img_x, $M_{perspective}^{(n)}$). The fitted perspective transformation may be combined with the location of the objects detected in the second sample images 214 (e.g., the bounding boxes). For each second sample image 214, the model training system 210 may transform, based on the fitted perspective transformation, the respective locations of the objects into respective transformed locations within the top plane view of the 3D space.

[0090] In some example embodiment, a location of an object detected from an image may be defined as a plurality of coordinates of a first plurality of key points of the object detected from the image. The key points for an object may include the four vertexes of a bounding box of the object. In some examples, the key points may further include a centroid of the object. Specifically, a location of an object in a second sample image 214

may be represented as coordinates of the four vertexes (left, top, right, and bottom) of a bounding box of the object within the second sample image 214, which may be represented as $(Co_{left}, Co_{top}, Co_{right}, Co_{bottom})$. In some example embodiments, a centroid of the object may also be taken into account to represent the location of the object. In some example embodiments, since the centroid point is not easy to locate and due to the absence of object information with 3D, the coordinates of the centroid may be represented as $Co_{c_x} = \frac{(Co_{left} + Co_{right})}{2}$, and $Co_{c_y} = Co_{bottom}$.

[0091] In some example embodiments, a key point transformation mapping may be determined for a second sample image 214 based on the fitted perspective transformation. The location of an object detected in a second sample image 214 may thus be represented as $(Co_{left}, Co_{top}, Co_{right}, Co_{bottom}, (Co_{c_x}, Co_{c_y}))$, i.e., five coordinates in the 2D space corresponding to the second sample image 214. The five coordinates may be split into five coordinate points, Kpt=[$(Co_{left}, Co_{top}), (Co_{right}, Co_{top}), (Co_{c_x}, Co_{c_y}), (Co_{left}, Co_{bottom}), (Co_{right}, Co_{bottom})$], corresponding to the coordinate points of top-left, top-right, center, bottom-left, bottom-right, where the center point is denoted as the center of the rigid object.

[0092] Further, the 2D coordinates of the location of the object may be filled with a predetermined value (e.g., 1.0) for the third dimensional, to transform $Kpt \in \mathbb{R}^{5\times2}$ to $Kpt \in \mathbb{R}^{5\times3}$. Then Kpt may be transformed based on the fitted perspective transformation, to obtain a transformed location Kpt' in the top plane view. The mapping may be represented as $Kpt' = Kpt \times M_{perspective}$. After scale normalization, the transformed location may be $Kpt' = \frac{Kpt'}{Kpt'[:,:,2]}$.

[0093] As a result, a second sample image 214 and its second label information 215 may be constructed as a training sample for the perspective transformation model 204 as $\{\text{image_x}, Kpt, Kpt', i, j, \text{p/n}\}$, where i and j represent a pair of objects in the second sample image 214 whose distance is to be estimated, the indicator "p" or "n" indicates whether the distance between the objects i and j is abnormal or not, Kpt indicates locations of the objects within the 2D space corresponding to the second sample image 214, and Kpt' indicates transformed locations of the objects in the 3D top plane view (which may be used to calculate the labeled distance between a pair of objects in the 3D top plane view).

[0094] In some example embodiments, if more than two pairs of objects of interest are detected from the second sample image 214, the objects i, j may be selected as the pair of object with the lowest transformed centroid distance, e.g., $Distancei,j = \underset{i,j}{argmin}(\left\| Kpt^{(i)} - Kpt^{(j)} \right\|)$ if the abnormal distance is determined as a distance lower than a distance threshold. For example, if a forklift and two pedestrians are detected in a second sample image 214 (as in the examples of FIG. 6A and FIG, 6B), then a pair of forklift and a pedestrian with the lowest distance may be selected for labeling the second sample image 214. In the case where an abnormal distance is determined as a distance exceeding a distance threshold, the pair of object with the highest transformed centroid distance may be selected. In some example embodiments, the second sample image 214 may be labeled with the second label information associated with more than one pair of objects whose distance(s) is to be estimated.

[0095] In some example embodiments, due to the large difference in the resolution of the scene camera, the real object size in the images corresponding to the aspect ratio of the sample images may have a large difference. Therefore, in the model training phase, it is considered to train the perspective transformation model 204 by varying the scale of the second sample image 214.

[0096] The resolution of a second sample image is represented as (h, w). To keep the spatial dimensions in the image from being further stretched in a destructive manner, the size of the second sample image may therefore be changed to $\text{Size} = (\text{h} / \max(\text{w}, \text{h}) * \text{IMAGE_SIZE}, \text{w} / \max(\text{w}, \text{h}) * \text{IMAGE_SIZE})$, where IMAGE_SIZE is set to the image scaling size.

[0097] At block 540, the model training system 210 constructs the perspective transformation model 204. The perspective transformation model 204 may be constructed with any suitable model structure in the computer vision (e.g., with MobileNet_v2 as the backbone model). The modeling of the perspective transformation model 204 may be represented as $Model(img) = (\widehat{M}_{perspective}, Scaler)$, with an image (represented as "img") as its input, and output a perspective transformation $\widehat{M}_{perspective}$ and a distance scale $Scaler$ for the input image.

[0098] At block 550, the model training system 210 trains the perspective transformation model 204 with the second sample images 214 and the second label

information 215. As mentioned above, for a specific perspective transformation model 204, a training sample may be constructed as follows: $\{\text{image_x}, Kpt, Kpt', i, j, \text{p/n}\}$, image_x represents a second sample image 214, and $(Kpt, Kpt', i, j, \text{p/n})$ represents the second label information 215.

[0099] In some example embodiments, the training of the perspective transformation model 204 may be based on a first loss function (also referred to as an objective function), which is used to evaluate a first error between a predicted indication of distance anomaly and a labeled indication of distance anomaly for a second sample image(s) 214. The first loss function may be represented as follows:

$$L_{dis} = \max\left(\left\|\widehat{Kpt}^{(i)} - \widehat{Kpt}^{(j)}\right\|_{l2} \cdot Scaler - a_{margin}\right) \cdot y_i,\ 0) \qquad (3)$$

[0100] The training objective is to decrease or minimize a loss value of the first loss function by iteratively updating or optimizing the perspective transformation model 204 (e.g., updating the model parameter values). The calculation of a loss value of the first loss function is as follows. For each second sample image 214, the model training system 210 may determine a predicted perspective transformation $\widehat{M}_{perspective}$ output by the perspective transformation model 204 for the second sample image 214 (based on the current model parameter values), and then determine respective predicted transformed locations of a pair of objects in the top plane view of the three-dimensional space using the predicted perspective transformation. A predicted transformed location of an object i is determined as $\widehat{Kpt}^{(i)}=Kpt^{(i)} \times \widehat{M}_{perspective}$, $Kpt^{(i)}$ is determined based on the location (coordinates) of the object i within the second sample image 214.

[0101] The model training system 210 may further determine a predicted distance between the pair of objects (e.g., objects i and j) by scaling a distance between the pair of transformed locations with a predicted distance scale output by the perspective transformation model for the second sample image. The predicted distance may be determined as $\left\|\widehat{Kpt}^{(i)} - \widehat{Kpt}^{(j)}\right\|_{l2} \cdot Scaler$, as in Equation (3).

[0102] The model training system 210 may determine a predicted indication of whether a distance between the pair of objects is abnormal based on a comparison result between the predicted distance and a distance threshold. For example, if a distance lower than the distance threshold is determined as an abnormal distance, the predicted distance lower

than the distance threshold may be indicated as an abnormal distance. Otherwise, if a distance exceeding a distance threshold is determined as an abnormal distance, the predicted distance exceeding the distance threshold may be indicated as an abnormal distance. The parameter a_{margin} in Equation (3) represents a distance threshold, and in this example, a distance lower than the distance threshold is determined as an abnormal distance.

[0103] In Equation (3), y_i is the labeled indication of whether a distance between objects i and j is abnormal or not, and y_i =1 indicates an abnormal distance, y_i = -1 indicates a normal distance. According to Equation (3), the loss value of the first loss function is lower or decreased as the predicted indication approximates to the labeled indication, which means that the perspective transformation model 204 has learned to predict relatively accurate perspective transformation and distance scale to make the correct anormal distance detection. Otherwise, the loss value of the first loss function is high or increased if the predicted indication is different from the labeled indication, which means that the predicted perspective transformation and the distance scale do not meet the expectation. The perspective transformation model 204 may need to be further optimized.

[0104] In some example embodiments, the training of the perspective transformation model 204 may be further based on a second loss function, which is used to evaluate a second error between a predicted transformed location(s) of an object(s) in a second sample image 214 and a labeled transformed location(s) of the object(s) in the second sample image 214. The second loss function may be represented as follows:

$$L_{kpt} = \sum_i \left| \widehat{Kpt}_i - Kpt'_i \right| \tag{4}$$

The training objective is to decrease or minimize a loss value of the second loss function by iteratively updating or optimizing the perspective transformation model 204 (e.g., updating the model parameter values). The calculation of a loss value of the first loss function is as follows. In Equation (3), for each second sample image 214, a predicted transformed location of an object i is determined as $\widehat{Kpt}^{(i)} = Kpt^{(i)} \times \hat{M}_{perspective}$, as described above. Kpt'_i indicates a labeled transformed location of the object i . $\widehat{Kpt}_i - Kpt'_i$ represents the difference (or error) between the predicted transformed location and the labeled transformed location. In some example embodiments, the predicted transformed locations and the labeled transformed locations for more objects detected in

the second sample image 214 (not only the objects whose distance is labeled for anomaly) may be determined, to calculate the loss value of the second loss function L_{kpt}.

[0105] In some example embodiments, the overall objective function for training the perspective transformation model 204 may be set to $L_{opt} = a \cdot L_{kpt} + (1.0 - a) \cdot L_{dis}$, where $a \in [0, 1]$ is used to control the weights of the two loss functions during the training process. The weight a may be initialized (e.g., to 0.5, 0.8, or the like) and may be adjusted or gradually reduced during the training (e.g., gradually reduced by 0.01 after a number of epochs of iterations). It would be appreciated that one or more additional or alternative loss functions may be designed for training the perspective transformation model 204 based on the second label information 215 for the second sample images 214.

[0106] In some example embodiments, the training of the perspective transformation model 204 may reference to the architecture 700 in FIG. 7. As illustrated, at Phase 1, the trained object detection model 202 is applied to detect objects from the second sample images 214, and thus the five coordinate points, Kpt=[$(Co_{left}, Co_{top}), (Co_{right}, Co_{top}), (Co_{c_x}, Co_{c_y}), (Co_{left}, Co_{bottom}), (Co_{right}, Co_{bottom})$] for each object may be derived based on the location of the detected object in the second sample images 214, for example, Kpt_i and Kpt_j. Through offline fitting of perspective transformation, the labeled transformed locations of the objects, e.g., Kpt'_i and Kpt'_j may be determined by mapping Kpt_i and Kpt_j with the fitted perspective transformation. At phase 2, the perspective transformation model 204 is trained, and the predicted transformed locations of the objects $\widehat{Kpt}$ may be determined and used to calculate the loss values of the first and/or second loss function.

[0107] In some example embodiments, because the model training stage involves segmented training process (the object detection model 202 and the perspective transformation model 204), considering the order of model optimization, the optimization accuracy of the bounding box of the object directly affects the accuracy of the subsequent calculation. Therefore, to ensure the accuracy of the perspective transformation models 204, in the process of model training, the generalization and robustness performance of the perspective model 204 may be further enhanced by randomly changing the locations for objects in the sample images. FIG. 8 illustrates an example 800 of floating key points in the model training. As illustrated, the locations for objects 811 and 812, Kpt_m and Kpt_n, in a sample image may be transformed into transformed locations Kpt'_m and Kpt'_n

in the top plane view 810. After the random changes, the locations for objects 811 and 812 may be adjusted to $\overline{Kpt_m}$ and $\overline{Kpt_n}$, and the transformed locations may also be changed as $\overline{Kpt'_m}$ and $\overline{Kpt'_n}$.

Object Distance Estimation and Abnormal Distance Detection in Inference Stage

[0108] FIG. 9 illustrates a flowchart of a process 900 for object distance estimation and abnormal distance detection in an inference stage in accordance with some example embodiments of the present disclosure. For the purpose of discussion, the process 900 will be described from the perspective of the model inference system 220 in FIG. 2.

[0109] At block 910, the model inference system 220 detects a first location of a first object and a second location of a second object within a 2D space corresponding to a target image 232 captured by a camera with a target deployment in a scene.

[0110] In some example embodiments, the model inference system 220 may detect the first location and the second location using the trained object detection model 202. The first object and the second object may be objects whose distance therebetween is to be estimated, e.g., for the anomaly detection or for other purposes.

[0111] In some example embodiments, through the detection, the first object or the second object may be detected in the target image 232 by respective bounding box. The first location of the first object may include a first plurality of coordinates of a first plurality of key points of the first object detected from the target image 232, and the second location may include a second plurality of coordinates of a second plurality of key points of the second object detected from the target image. The first plurality of key points may at least include a first centroid of the first object, and the second plurality of key points may at least include a second centroid of the second object.

[0112] The key points for an object may include coordinates of the four vertexes of a bounding box of the object and the centroid of the object, which may be represented as $(Co_{left}, Co_{top}, Co_{right}, Co_{bottom}, Co_{c_x} = \frac{(Co_{left} + Co_{right})}{2}, Co_{c_y} = Co_{bottom})$. In some examples, the five coordinates may be split into five coordinate points to represent, Kpt=[$(Co_{left}, Co_{top}), (Co_{right}, Co_{top}), (Co_{c_x}, Co_{c_y}), (Co_{left}, Co_{bottom}), (Co_{right}, Co_{bottom})$], to represent a location of the corresponding location.

[0113] At block 920, the model inference system 220 determines a perspective

transformation $\widehat{M}_{perspective}$ and a distance scale $Scaler$ for the target image 232 by providing the target image 232 into a trained perspective transformation model 204 associated with the target deployment. The perspective transformation model 204 used for the target image 232 may be selected from a plurality of perspective transformation models 204 based on a target deployment identity of the target deployment. The plurality of perspective transformation models 204 are each identified with respective deployment identities. The deployment identity may sometimes correspond to a scene ID for the scene where the camera is deployed.

[0114] At block 930, the model inference system 220 transforms, based on the determined perspective transformation $\widehat{M}_{perspective}$, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively. For example, a transformed location for an object may be determined as $Kpt' = Kpt \times \widehat{M}_{perspective}$.

[0115] At block 940, the model inference system 220 determines a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale $Scaler$.

[0116] In some example, the first transformed location and the second transformed location may be normalized, though $Kpt' = \frac{Kpt'}{Kpt'[:,:,2]}$. In some example embodiments, the target distance between the first transformed location and the second transformed location may be calculated as $distance = \left\| Kpt'^{(i)} - Kpt'^{(j)} \right\|_{l2} \cdot Scaler$. In some example embodiments, the coordinates of the centroids of the first and second objects may be used to determine the distance between two objects.

[0117] In some example embodiments, the model inference system 220 may further determine an indication of whether the target distance between the first object and the second object is abnormal based on a comparison result between the target distance and a distance threshold. The distance threshold may be configured according to the actual applications of distance anomaly detection. In some example embodiments, if a distance lower than the distance threshold is considered as abnormal, the model inference system 220 may determine whether the target distance is abnormal or not based on a

determination of whether the target distance is lower than the distance threshold. In some example embodiments, if a distance exceeding the distance threshold is considered as abnormal, the model inference system 220 may determine whether the target distance is abnormal or not based on a determination of whether the target distance exceeds the distance threshold.

[0118] In some example embodiments, actions may be determined and performed according to the distance anomaly detection. As some examples, if it is determined that the distance(s) between objects is abnormal, alarms may be generated and transmitted to one or more related devices, and/or operation commands may be sent to the related devices. In the example use case of detecting whether a vehicle or forklift is too close to a pedestrian(s), in response to a detection of the abnormal distance between the vehicle or forklift, an operation command of applying a force stop may be sent to the vehicle or forklift, and/or an alarming device may be activated to warn the driver and the pedestrian(s) of the dangerous situation in the scene.

[0119] The present disclosure provides solutions for object distance estimation and distance anomaly detection based on monocular image data, which provides a lightweight detection method for video distance anomaly monitoring application scenarios, and effectively reduces the implementation cost and usage cost. According to the example embodiments of the present disclosure, instead of performing a complete 2D-3D scene restoration to achieve the estimation of the distance between objects, the models are trained and applied to transform pixel location in an image into locations in the 3D top plane view, to calculate the distance between the objects is a dangerous distance. This can avoid a lot of tedious information collection and feature construction work in the early stage, and reduce the complexity of the model and its absolute requirements for data accuracy.

[0120] Further, in the training of the models, there is no need for the precise distance labeling for the 3D scenes in absolute sense. Thus, from the perspective of modelling, the example embodiments of the present disclosure may not need to strictly define the absolute coordinate positions fitted in the 3D scenes. Due to the full consideration of the fitting of perspective transformation and the fitting of abnormal distance, the distance estimation and the anomaly detection based on the model output are still reliable.

Example Apparatus, Device and Medium

[0121] The present application may be described in the general context of computer-executable instructions executed by a computer, such as a program module or unit. Generally, a program module or unit may include routines, programs, objects, components, data structures, and the like that perform a particular task to implement a particular abstract data type. In general, the program module or unit may be implemented by software, hardware, or a combination of both. The present application may also be practiced in distributed computing environments where the program modules or units may be located in local and process computer storage media, including storage devices, or by storage in portable computing devices for abnormally distant data transfer.

[0122] In some example embodiments, an apparatus capable of performing any of the processes 300, 400, 500, and 900 may comprise means for performing the respective operations of those processes. The means may be implemented in any suitable form. For example, the means may be implemented in a circuitry or software module. The first apparatus may be implemented as or included in the model inference system 220 and/or the model training system 210 in FIG. 2.

[0123] In some example embodiments, the apparatus comprises means for means for detecting a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene; means for determining a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment; means for transforming, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and means for determining a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

[0124] In some example embodiments, the first apparatus further comprises: means for determining an indication of whether the target distance between the first object and the second object is abnormal based on a comparison result between the target distance and a distance threshold.

[0125] In some example embodiments, the means for determining the perspective

transformation and the distance scale for the target image comprises: means for selecting the perspective transformation model associated with the target deployment from a plurality of perspective transformation models based on a target deployment identity of the target deployment, the plurality of perspective transformation models being identified with respective deployment identities.

[0126] In some example embodiments, the target deployment is identified based on a mounting height, a shooting orientation, and a type of the target camera.

[0127] In some example embodiments, the means for detecting the first location and the second location comprises: means for detecting the first location and the second location using an object detection model, wherein the object detection model is trained with first sample images captured by a plurality of cameras with a plurality of deployments and first label information indicating objects in the first sample images.

[0128] In some example embodiments, the first location comprises a first plurality of coordinates of a first plurality of key points of the first object detected from the target image, and the second location comprises a second plurality of coordinates of a second plurality of key points of the second object detected from the target image, wherein the first plurality of key points at least comprises a first centroid of the first object, and the second plurality of key points at least comprises a second centroid of the second object.

[0129] In some example embodiments, the perspective transformation model is trained using second sample images captured by a camera with the target deployment and second label information indicating at least one of the following for each second sample image: a labeled indication of whether a distance between the pair of objects is abnormal, the pair of objects being detected in the second sample image, or respective labeled transformed locations of the pair of objects in the top plane view of the three-dimensional space.

[0130] In some example embodiments, the perspective transformation model is trained by the following: determining respective predicted transformed locations of the pair of objects in the top plane view of the three-dimensional space using a predicted perspective transformation output by the perspective transformation model for the second sample image; determining a predicted distance between the pair of objects by scaling a distance between the pair of transformed locations with a predicted distance scale output by the perspective transformation model for the second sample image; determine a predicted indication of whether a distance between the pair of objects is abnormal based on a

comparison result between the predicted distance and a distance threshold; determining a first loss value of a first loss function based on a first error between the predicted indication and the labeled indication for the second sample image; and updating the perspective transformation model by decreasing the first loss value of the first loss function.

[0131] In some example embodiments, the perspective transformation model is further trained by the following: determining a second loss value of a second loss function based on second errors between the respective predicted transformed locations and the respective labeled transformed locations; and updating the perspective transformation model by decreasing the second loss value of the second loss function.

[0132] In some example embodiments, the respective labeled transformed locations of the pair of objects are determined by the following: determining a fitted perspective transformation associated with the target deployment based on respective locations of the pair of objects within a two-dimensional space corresponding to the second sample image; transforming, based on the fitted perspective transformation, the respective locations of the pair of objects into the respective transformed locations within the top plane view of the three-dimensional space.

[0133] In some example embodiments, the target camera is a monocular camera.

[0134] In some example embodiments, the apparatus further comprises means for performing other operations in some example embodiments of the present disclosure. In some example embodiments, the means comprises at least one processor; and at least one memory storing instructions that, when executed by the at least one processor, cause the performance of the apparatus.

[0135] FIG. 10 is a simplified block diagram of a device 1000 that is suitable for implementing example embodiments of the present disclosure. The device 1000 may be provided to implement a communication device, for example, the model inference system 220 and/or the model training system 210 as shown in FIG. 2. As shown, the device 1000 includes one or more processors 1010, one or more memories 1020 coupled to the processor 1010, and one or more communication modules 1040 coupled to the processor 1010.

[0136] The communication module 1040 is for bidirectional communications. The

communication module 1040 has one or more communication interfaces to facilitate communication with one or more other modules or devices. The communication interfaces may represent any interface that is necessary for communication with other network elements. In some example embodiments, the communication module 1040 may include at least one antenna.

[0137] The processor 1010 may be of any type suitable to the local technical network and may include one or more of the following: general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multicore processor architecture, as non-limiting examples. The device 1000 may have multiple processors, such as an application specific integrated circuit chip that is slaved in time to a clock which synchronizes the main processor.

[0138] The memory 1020 may include one or more non-volatile memories and one or more volatile memories. Examples of the non-volatile memories include, but are not limited to, a Read Only Memory (ROM) 1024, an electrically programmable read only memory (EPROM), a flash memory, a hard disk, a compact disc (CD), a digital video disk (DVD), an optical disk, a laser disk, and other magnetic storage and/or optical storage. Examples of the volatile memories include, but are not limited to, a random access memory (RAM) 1022 and other volatile memories that will not last in the power-down duration.

[0139] A computer program 1030 includes computer executable instructions that are executed by the associated processor 1010. The instructions of the program 1030 may include instructions for performing operations/acts of some example embodiments of the present disclosure. The program 1030 may be stored in the memory, e.g., the ROM 1024. The processor 1010 may perform any suitable actions and processing by loading the program 1030 into the RAM 1022.

[0140] The example embodiments of the present disclosure may be implemented by means of the program 1030 so that the device 1000 may perform any process of the disclosure as discussed above. The example embodiments of the present disclosure may also be implemented by hardware or by a combination of software and hardware.

[0141] In some example embodiments, the program 1030 may be tangibly contained in a computer readable medium which may be included in the device 1000 (such as in the memory 1020) or other storage devices that are accessible by the device 1000. The device

1000 may load the program 1030 from the computer readable medium to the RAM 1022 for execution. In some example embodiments, the computer readable medium may include any types of non-transitory storage medium, such as ROM, EPROM, a flash memory, a hard disk, CD, DVD, and the like. The term "non-transitory," as used herein, is a limitation of the medium itself (i.e., tangible, not a signal) as opposed to a limitation on data storage persistency (e.g., RAM vs. ROM).

[0142] FIG. 11 shows an example of the computer readable medium 1100 which may be in form of CD, DVD or other optical storage disk. The computer readable medium 1100 has the program 1030 stored thereon.

[0143] Generally, various embodiments of the present disclosure may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. Some aspects may be implemented in hardware, and other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device. Although various aspects of embodiments of the present disclosure are illustrated and described as block diagrams, flowcharts, or using some other pictorial representations, it is to be understood that the block, apparatus, system, technique or method described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[0144] Some example embodiments of the present disclosure also provide at least one computer program product tangibly stored on a computer readable medium, such as a non-transitory computer readable medium. The computer program product includes computer-executable instructions, such as those included in program modules, being executed in a device on a target physical or virtual processor, to carry out any of the methods as described above. Generally, program modules include routines, programs, libraries, objects, classes, components, data structures, or the like that perform particular tasks or implement particular abstract data types. The functionality of the program modules may be combined or split between program modules as desired in various embodiments. Machine-executable instructions for program modules may be executed within a local or distributed device. In a distributed device, program modules may be located in both local and remote storage media.

[0145] Program code for carrying out methods of the present disclosure may be written

in any combination of one or more programming languages. The program code may be provided to a processor or controller of a general purpose computer, special purpose computer, or other programmable data processing apparatus, such that the program code, when executed by the processor or controller, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may execute entirely on a machine, partly on the machine, as a stand-alone software package, partly on the machine and partly on a remote machine or entirely on the remote machine or server.

[0146] In the context of the present disclosure, the computer program code or related data may be carried by any suitable carrier to enable the device, apparatus or processor to perform various processes and operations as described above. Examples of the carrier include a signal, computer readable medium, and the like.

[0147] The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable medium may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the computer readable storage medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[0148] Further, although operations are depicted in a particular order, this should not be understood as requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results. In certain circumstances, multitasking and parallel processing may be advantageous. Likewise, although several specific implementation details are contained in the above discussions, these should not be construed as limitations on the scope of the present disclosure, but rather as descriptions of features that may be specific to particular embodiments. Unless explicitly stated, certain features that are described in the context of separate embodiments may also be implemented in combination in a single embodiment. Conversely, unless explicitly stated, various features that are described in the context of a single embodiment may also be implemented in a plurality of embodiments separately or in any suitable sub-combination.

[0149] Although the present disclosure has been described in languages specific to structural features and/or methodological acts, it is to be understood that the present disclosure defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts described above are disclosed as example forms of implementing the claims.

WHAT IS CLAIMED IS:

1. An apparatus comprising:

at least one processor; and

at least one memory storing instructions that, when executed by the at least one processor, cause the apparatus at least to:

detect a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene;

determine a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment;

transform, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and

determine a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

2. The apparatus of claim 1, wherein the at least one memory and the at least one processor further cause the apparatus to:

determine an indication of whether the target distance between the first object and the second object is abnormal based on a comparison result between the target distance and a distance threshold.

3. The apparatus of claim 1 or 2, wherein the at least one memory and the at least one processor cause the apparatus to determine the perspective transformation and the distance scale for the target image by:

selecting the perspective transformation model associated with the target deployment from a plurality of perspective transformation models based on a target deployment identity of the target deployment, the plurality of perspective transformation models being identified with respective deployment identities.

4. The apparatus of claim 3, wherein the target deployment is identified based on a mounting height, a shooting orientation, and a type of the target camera.

5. The apparatus of any of claims 1 to 4, wherein the at least one memory and the at least one processor cause the apparatus to detect the first location and the second location by:

detecting the first location and the second location using an object detection model,

wherein the object detection model is trained with first sample images captured by a plurality of cameras with a plurality of deployments and first label information indicating objects in the first sample images.

6. The apparatus of any of claims 1 to 5, wherein the first location comprises a first plurality of coordinates of a first plurality of key points of the first object detected from the target image, and the second location comprises a second plurality of coordinates of a second plurality of key points of the second object detected from the target image, wherein the first plurality of key points at least comprises a first centroid of the first object, and the second plurality of key points at least comprises a second centroid of the second object.

7. The apparatus of any of claims 1 to 6, wherein the perspective transformation model is trained using second sample images captured by a camera with the target deployment and second label information indicating at least one of the following for each second sample image:

a labeled indication of whether a distance between the pair of objects is abnormal, the pair of objects being detected in the second sample image, or

respective labeled transformed locations of the pair of objects in the top plane view of the three-dimensional space.

8. The apparatus of claim 7, wherein the perspective transformation model is trained by the following:

determining respective predicted transformed locations of the pair of objects in the top plane view of the three-dimensional space using a predicted perspective transformation output by the perspective transformation model for the second sample image;

determining a predicted distance between the pair of objects by scaling a distance between the pair of transformed locations with a predicted distance scale output by the perspective transformation model for the second sample image;

determine a predicted indication of whether a distance between the pair of objects is abnormal based on a comparison result between the predicted distance and a distance threshold;

determining a first loss value of a first loss function based on a first error between the predicted indication and the labeled indication for the second sample image; and

updating the perspective transformation model by decreasing the first loss value of the first loss function.

9. The apparatus of claim 7, wherein the perspective transformation model is further trained by the following:

determining a second loss value of a second loss function based on second errors between the respective predicted transformed locations and the respective labeled transformed locations; and

updating the perspective transformation model by decreasing the second loss value of the second loss function.

10. The apparatus of claim 7, wherein the respective labeled transformed locations of the pair of objects are determined by the following:

determining a fitted perspective transformation associated with the target deployment based on respective locations of the pair of objects within a two-dimensional space corresponding to the second sample image;

transforming, based on the fitted perspective transformation, the respective locations of the pair of objects into the respective transformed locations within the top plane view of the three-dimensional space.

11. The apparatus of any of claims 1 to 10, wherein the target camera is a monocular camera.

12. A method comprising:

detecting a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene;

determining a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment;

transforming, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and

determining a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

13. The method of claim 12, further comprising:

determining an indication of whether the target distance between the first object and the second object is abnormal based on a comparison result between the target distance and a distance threshold.

14. The method of claim 12 or 13, wherein determining the perspective transformation and the distance scale for the target image comprises:

selecting the perspective transformation model associated with the target deployment from a plurality of perspective transformation models based on a target deployment identity of the target deployment, the plurality of perspective transformation models being identified with respective deployment identities.

15. The method of claim 14, wherein the target deployment is identified based on a mounting height, a shooting orientation, and a type of the target camera.

16. The method of any of claims 12 to 15, wherein detecting the first location and the second location comprises:

detecting the first location and the second location using an object detection model,

wherein the object detection model is trained with first sample images captured by a plurality of cameras with a plurality of deployments and first label information indicating objects in the first sample images.

17. The method of any of claims 12 to 16, wherein the first location comprises a first plurality of coordinates of a first plurality of key points of the first object detected from the target image, and the second location comprises a second plurality of coordinates of a second plurality of key points of the second object detected from the target image, wherein the first

plurality of key points at least comprises a first centroid of the first object, and the second plurality of key points at least comprises a second centroid of the second object.

18. The method of any of claims 12 to 17, wherein the perspective transformation model is trained using second sample images captured by a camera with the target deployment and second label information indicating at least one of the following for each second sample image:

a labeled indication of whether a distance between the pair of objects is abnormal, the pair of objects being detected in the second sample image, or

respective labeled transformed locations of the pair of objects in the top plane view of the three-dimensional space.

19. The method of claim 18, wherein the perspective transformation model is trained by the following:

determining respective predicted transformed locations of the pair of objects in the top plane view of the three-dimensional space using a predicted perspective transformation output by the perspective transformation model for the second sample image;

determining a predicted distance between the pair of objects by scaling a distance between the pair of transformed locations with a predicted distance scale output by the perspective transformation model for the second sample image;

determine a predicted indication of whether a distance between the pair of objects is abnormal based on a comparison result between the predicted distance and a distance threshold;

determining a first loss value of a first loss function based on a first error between the predicted indication and the labeled indication for the second sample image; and

updating the perspective transformation model by decreasing the first loss value of the first loss function.

20. The method of claim 18, wherein the perspective transformation model is further trained by the following:

determining a second loss value of a second loss function based on second errors between the respective predicted transformed locations and the respective labeled transformed locations; and

updating the perspective transformation model by decreasing the second loss value of the second loss function.

21. The method of claim 18, wherein the respective labeled transformed locations of the pair of objects are determined by the following:

determining a fitted perspective transformation associated with the target deployment based on respective locations of the pair of objects within a two-dimensional space corresponding to the second sample image;

transforming, based on the fitted perspective transformation, the respective locations of the pair of objects into the respective transformed locations within the top plane view of the three-dimensional space.

22. An apparatus comprising:

means for detecting a first location of a first object and a second location of a second object within a two-dimensional space corresponding to a target image captured by a camera with a target deployment in a scene;

means for determining a perspective transformation and a distance scale for the target image by providing the target image into a perspective transformation model associated with the target deployment;

means for transforming, based on the determined perspective transformation, the first location and the second location into a first transformed location of the first object and a second transformed location of the second object within a top plane view of a three-dimensional space, respectively; and

means for determining a target distance between the first object and the second object in a three-dimensional space by scaling a distance between the first transformed location and the second transformed location with the distance scale.

23. A computer readable medium comprising instructions which, when executed by an apparatus, cause the apparatus to perform at least the method of any of claims 12 to 21.

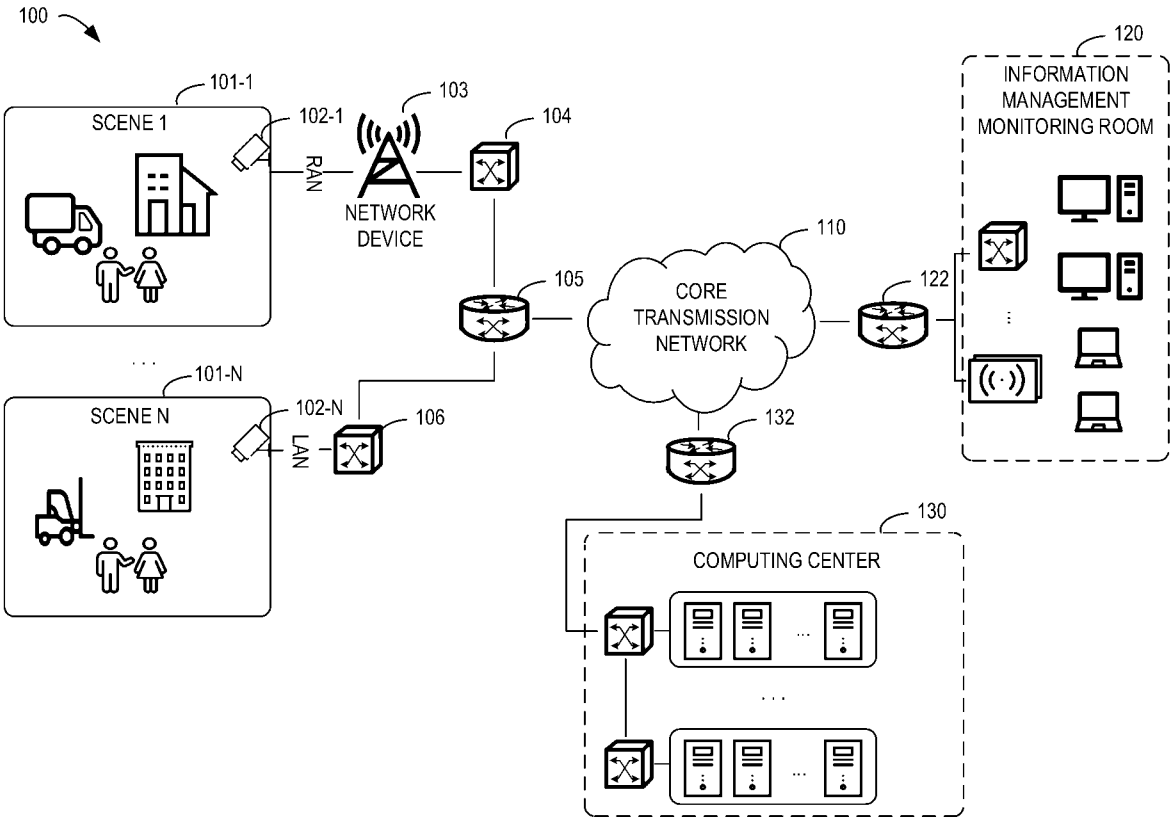

100
101-1
SCENE 1
102-1
RAN
103
NETWORK DEVICE
104
105
110
CORE TRANSMISSION NETWORK
122
120
INFORMATION MANAGEMENT MONITORING ROOM
101-N
SCENE N
102-N
LAN
106
132
130
COMPUTING CENTER

FIG. 1

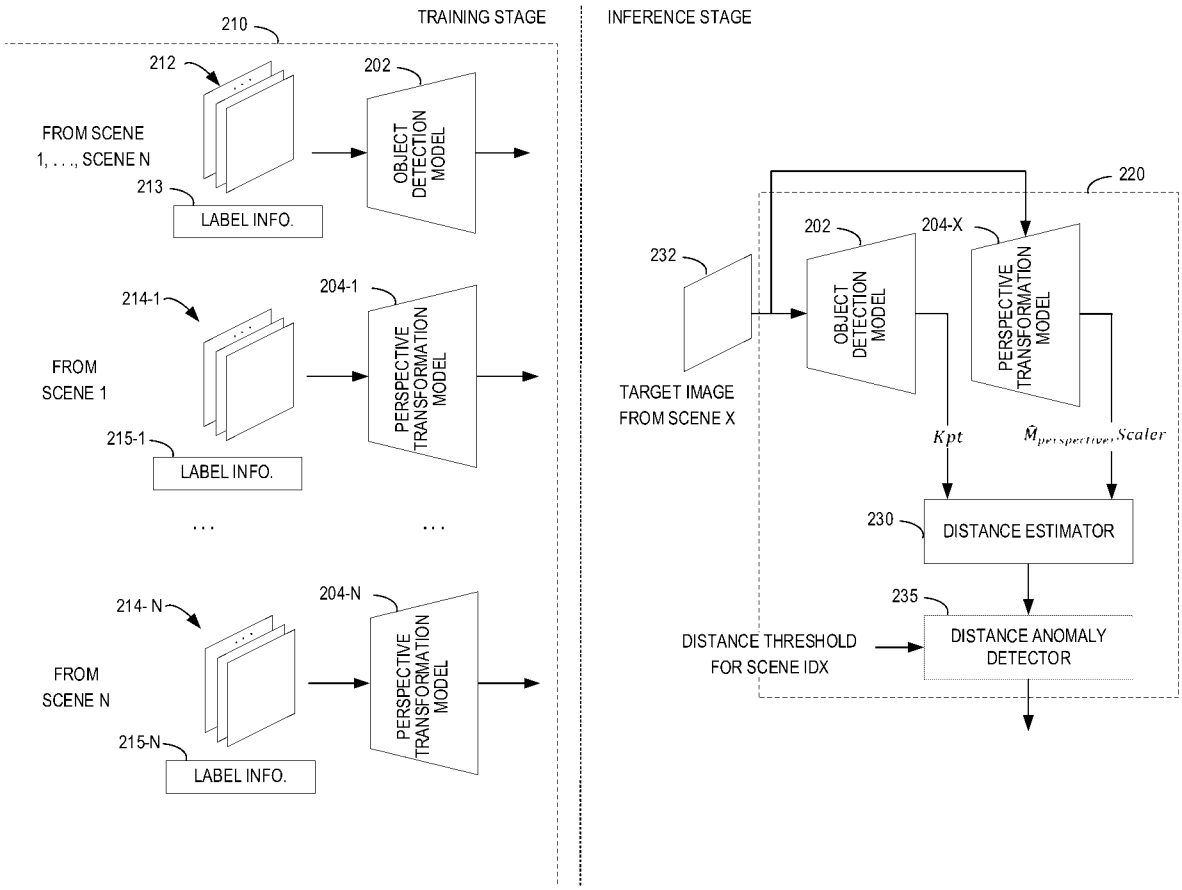

TRAINING STAGE
INFERENCE STAGE
210
212
202
FROM SCENE 1, . . ., SCENE N
213
LABEL INFO.
OBJECT DETECTION MODEL
214-1
204-1
FROM SCENE 1
215-1
LABEL INFO.
PERSPECTIVE TRANSFORMATION MODEL
214- N
204-N
FROM SCENE N
215-N
LABEL INFO.
220
232
202
204-X
TARGET IMAGE FROM SCENE X
OBJECT DETECTION MODEL
PERSPECTIVE TRANSFORMATION MODEL
Kpt
$\hat{M}_{perspective}$
Scaler
230
DISTANCE ESTIMATOR
235
DISTANCE THRESHOLD FOR SCENE IDX
DISTANCE ANOMALY DETECTOR

FIG. 2

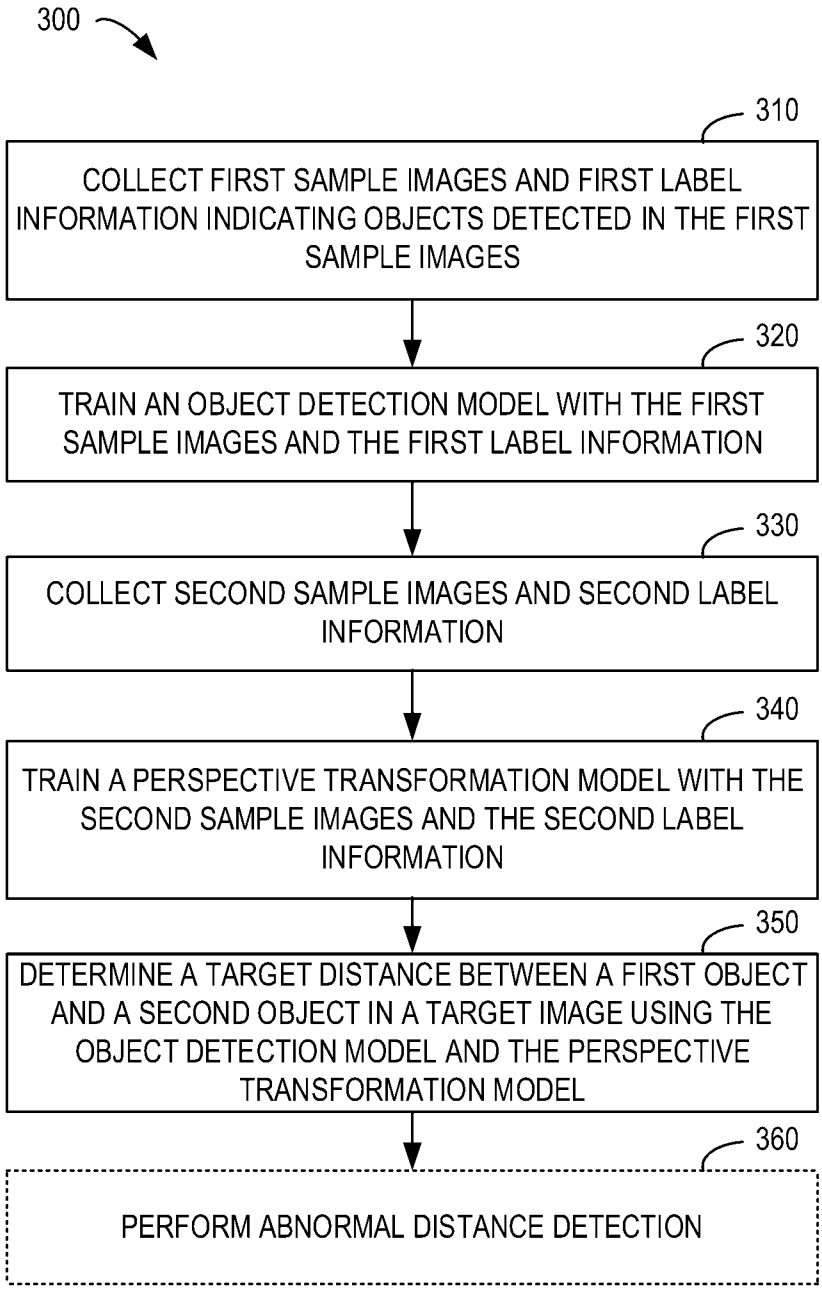

300
310
COLLECT FIRST SAMPLE IMAGES AND FIRST LABEL INFORMATION INDICATING OBJECTS DETECTED IN THE FIRST SAMPLE IMAGES
320
TRAIN AN OBJECT DETECTION MODEL WITH THE FIRST SAMPLE IMAGES AND THE FIRST LABEL INFORMATION
330
COLLECT SECOND SAMPLE IMAGES AND SECOND LABEL INFORMATION
340
TRAIN A PERSPECTIVE TRANSFORMATION MODEL WITH THE SECOND SAMPLE IMAGES AND THE SECOND LABEL INFORMATION
350
DETERMINE A TARGET DISTANCE BETWEEN A FIRST OBJECT AND A SECOND OBJECT IN A TARGET IMAGE USING THE OBJECT DETECTION MODEL AND THE PERSPECTIVE TRANSFORMATION MODEL
360
PERFORM ABNORMAL DISTANCE DETECTION

FIG. 3

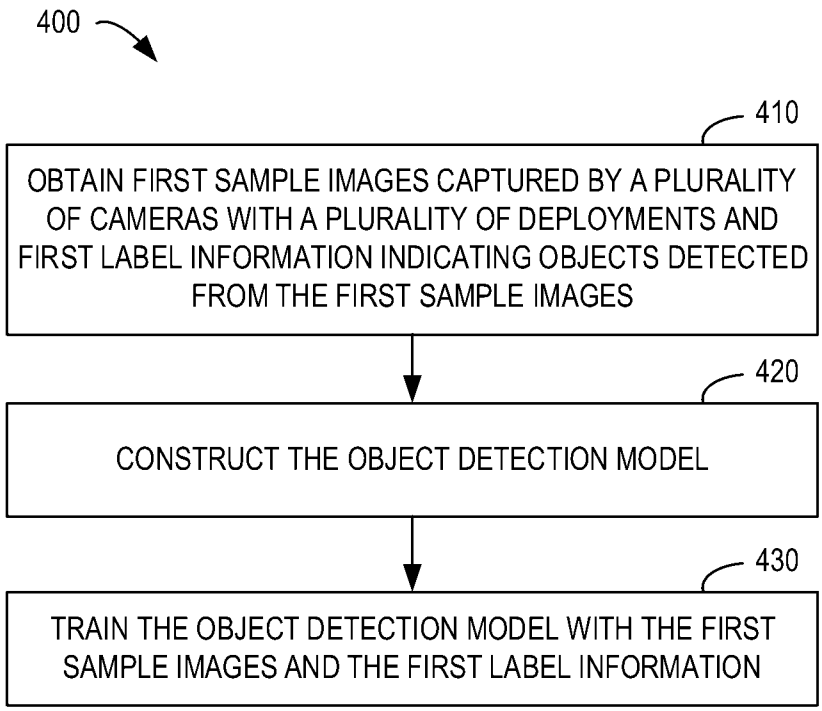

400
410
OBTAIN FIRST SAMPLE IMAGES CAPTURED BY A PLURALITY OF CAMERAS WITH A PLURALITY OF DEPLOYMENTS AND FIRST LABEL INFORMATION INDICATING OBJECTS DETECTED FROM THE FIRST SAMPLE IMAGES
420
CONSTRUCT THE OBJECT DETECTION MODEL
430
TRAIN THE OBJECT DETECTION MODEL WITH THE FIRST SAMPLE IMAGES AND THE FIRST LABEL INFORMATION

FIG. 4

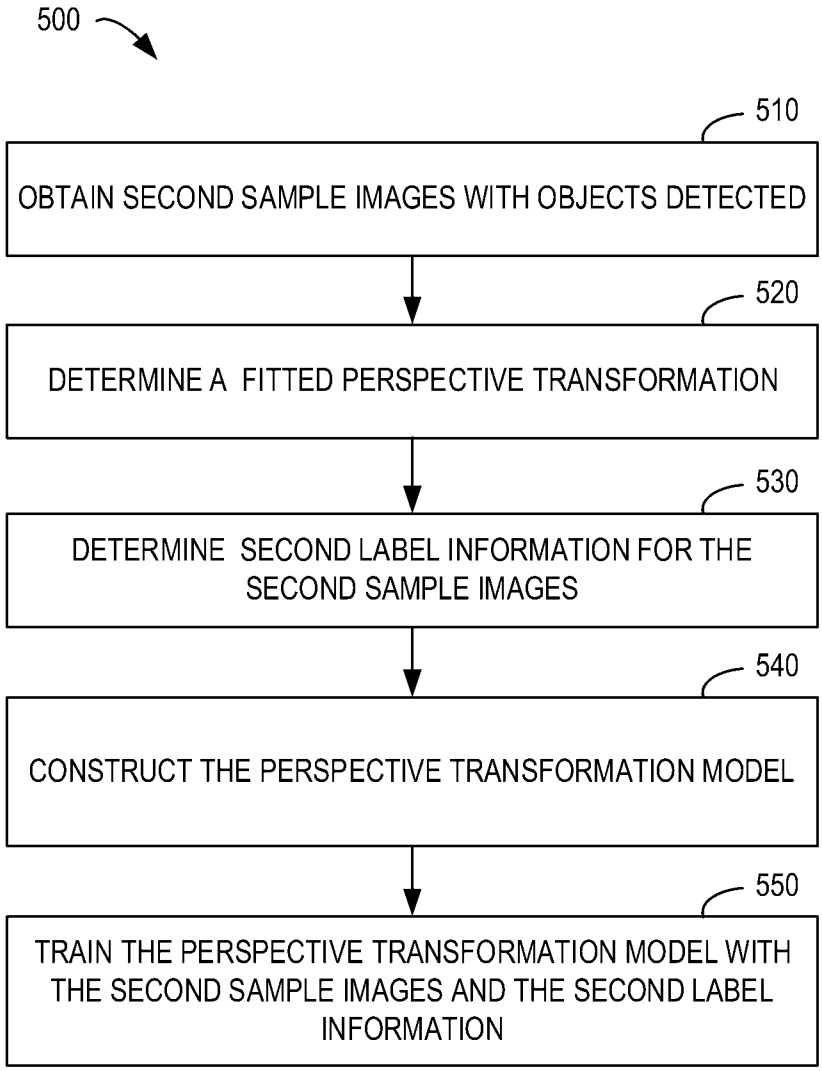

500
510
OBTAIN SECOND SAMPLE IMAGES WITH OBJECTS DETECTED
520
DETERMINE A FITTED PERSPECTIVE TRANSFORMATION
530
DETERMINE SECOND LABEL INFORMATION FOR THE SECOND SAMPLE IMAGES
540
CONSTRUCT THE PERSPECTIVE TRANSFORMATION MODEL
550
TRAIN THE PERSPECTIVE TRANSFORMATION MODEL WITH THE SECOND SAMPLE IMAGES AND THE SECOND LABEL INFORMATION

FIG. 5

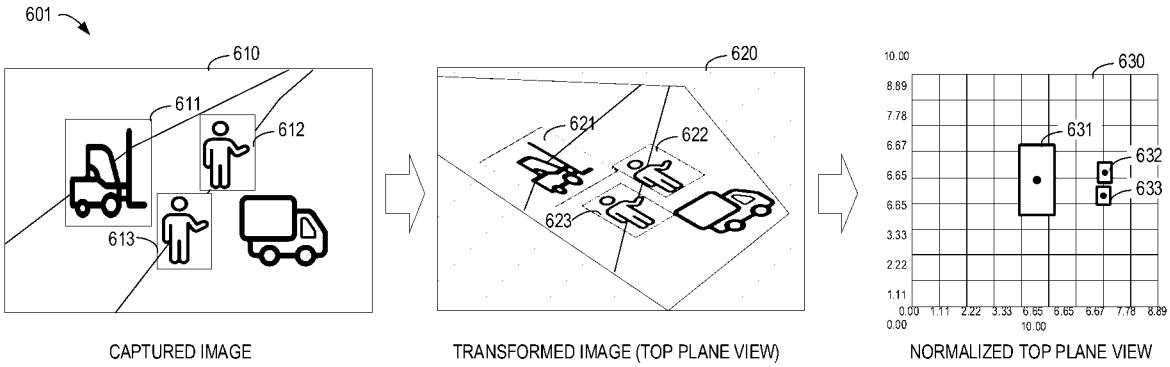

601
610
611
612
613
620
621
622
623
630
631
632
633
10.00
8.89
7.78
6.67
6.65
6.65
3.33
2.22
1.11
0.00
1.11
2.22
3.33
6.65
6.65
6.67
7.78
8.89
10.00
CAPTURED IMAGE
TRANSFORMED IMAGE (TOP PLANE VIEW)
NORMALIZED TOP PLANE VIEW

FIG. 6A

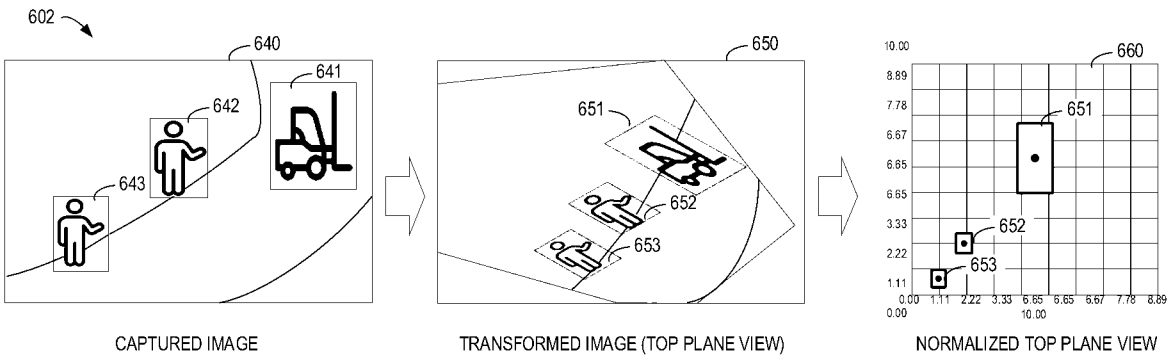

602
640
641
642
643
650
651
652
653
660
651
652
653
10.00
8.89
7.78
6.67
6.65
6.65
3.33
2.22
1.11
0.00
1.11
2.22
3.33
6.65
6.65
6.67
7.78
8.89
10.00
CAPTURED IMAGE
TRANSFORMED IMAGE (TOP PLANE VIEW)
NORMALIZED TOP PLANE VIEW

FIG. 6B

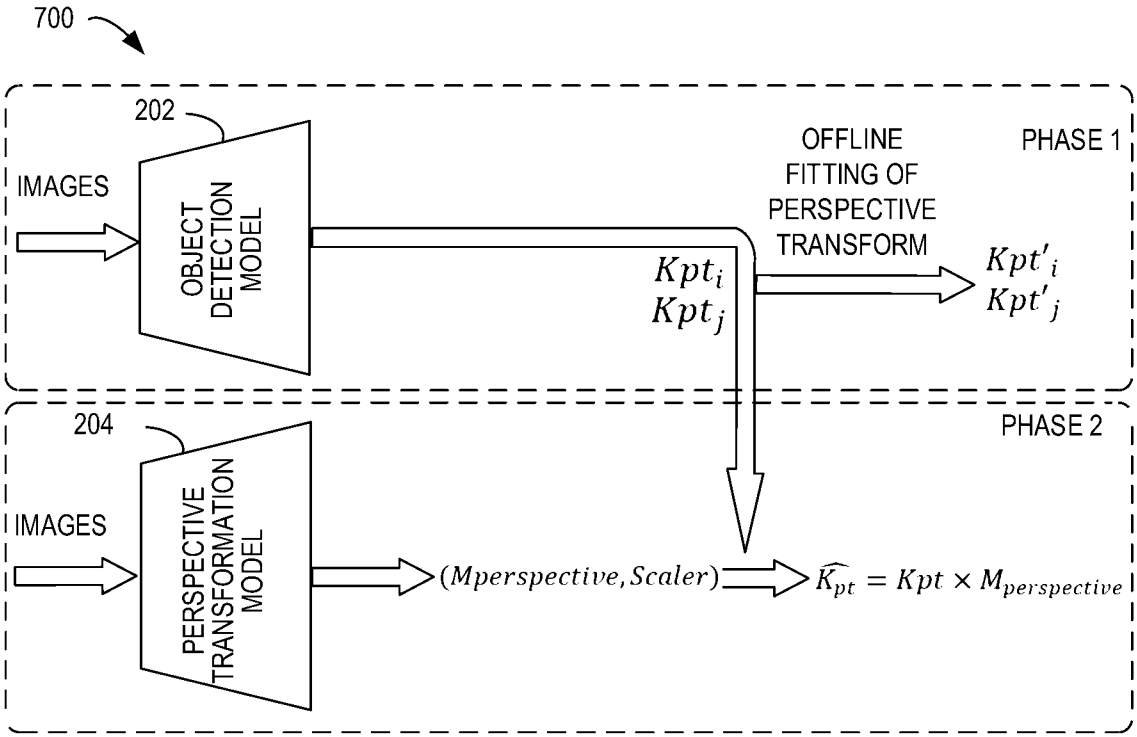

700
202
IMAGES
OBJECT DETECTION MODEL
OFFLINE FITTING OF PERSPECTIVE TRANSFORM
PHASE 1
Kpt_i
Kpt_j
Kpt'_i
Kpt'_j
204
PHASE 2
IMAGES
PERSPECTIVE TRANSFORMATION MODEL
$(Mperspective, Scaler)$
$\widehat{K_{pt}} = Kpt \times M_{perspective}$

FIG. 7

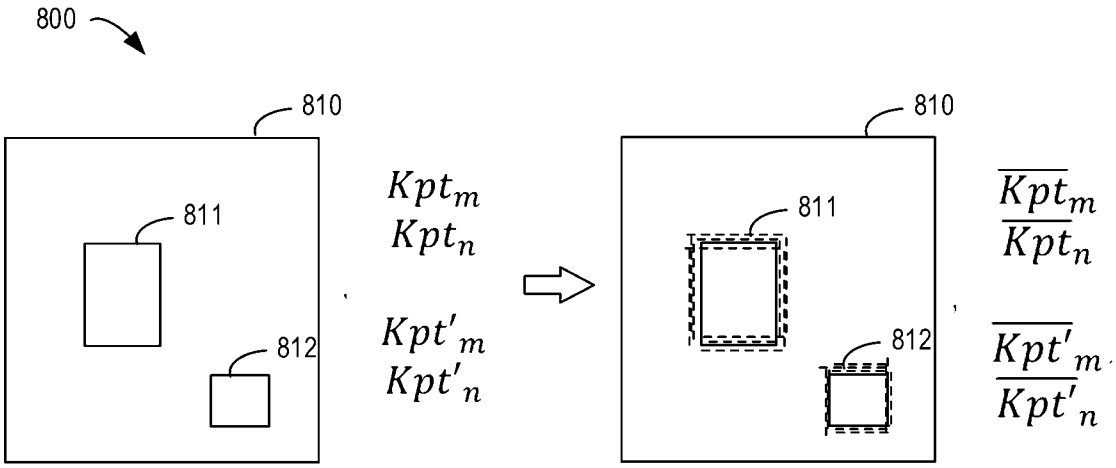

800
810
811
812
Kpt_m
Kpt_n
Kpt'_m
Kpt'_n
810
811
812
$\overline{Kpt_m}$
$\overline{Kpt_n}$
$\overline{Kpt'_m}$
$\overline{Kpt'_n}$

FIG. 8

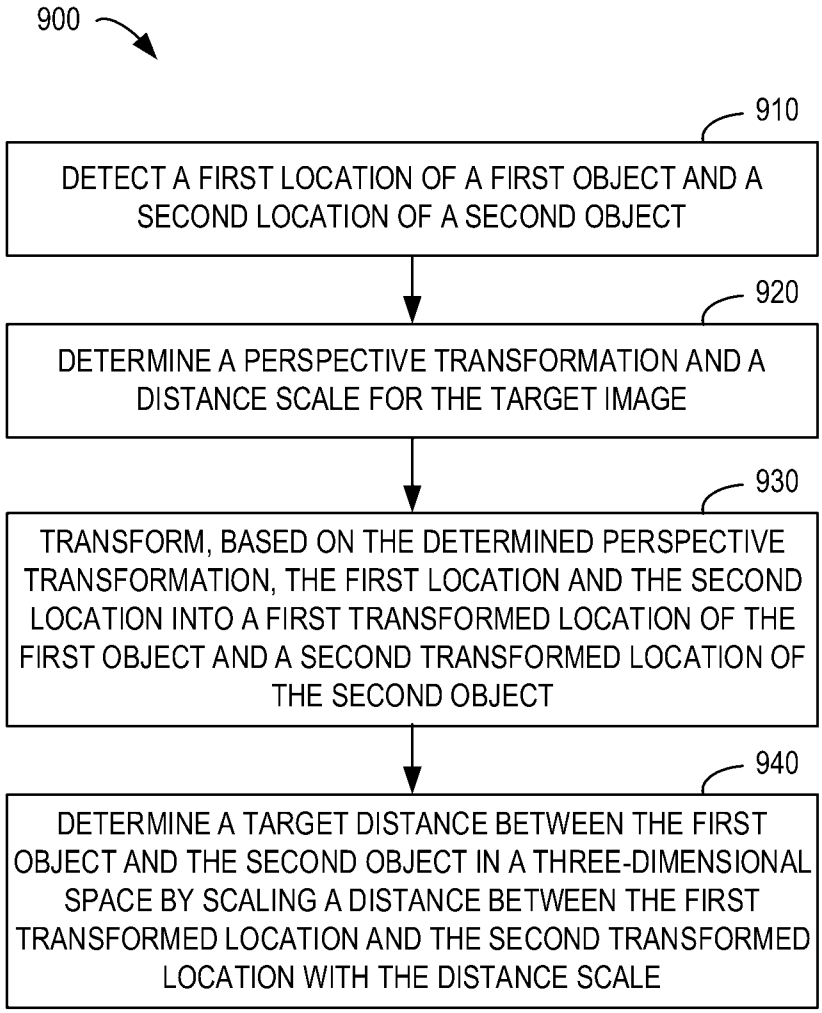

900
910
DETECT A FIRST LOCATION OF A FIRST OBJECT AND A SECOND LOCATION OF A SECOND OBJECT
920
DETERMINE A PERSPECTIVE TRANSFORMATION AND A DISTANCE SCALE FOR THE TARGET IMAGE
930
TRANSFORM, BASED ON THE DETERMINED PERSPECTIVE TRANSFORMATION, THE FIRST LOCATION AND THE SECOND LOCATION INTO A FIRST TRANSFORMED LOCATION OF THE FIRST OBJECT AND A SECOND TRANSFORMED LOCATION OF THE SECOND OBJECT
940
DETERMINE A TARGET DISTANCE BETWEEN THE FIRST OBJECT AND THE SECOND OBJECT IN A THREE-DIMENSIONAL SPACE BY SCALING A DISTANCE BETWEEN THE FIRST TRANSFORMED LOCATION AND THE SECOND TRANSFORMED LOCATION WITH THE DISTANCE SCALE

FIG. 9

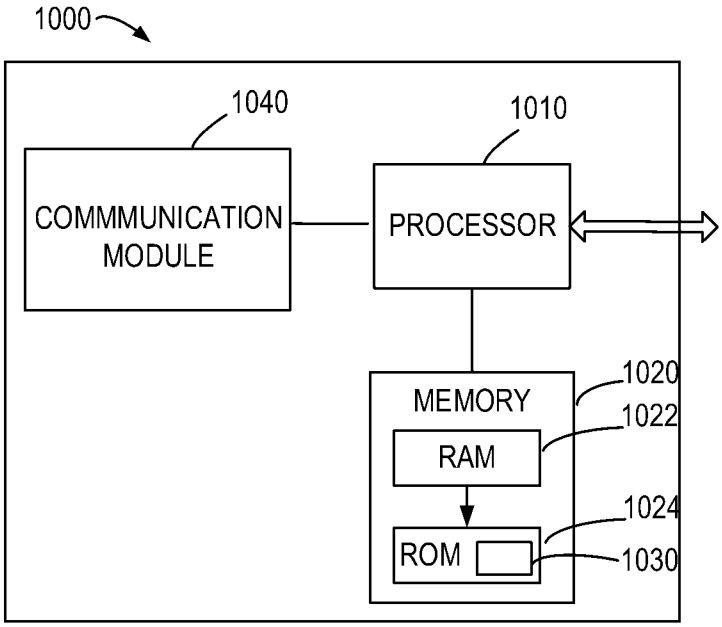

1000
1040
1010
COMMMUNICATION MODULE
PROCESSOR
MEMORY
1020
1022
RAM
1024
ROM
1030

FIG. 10

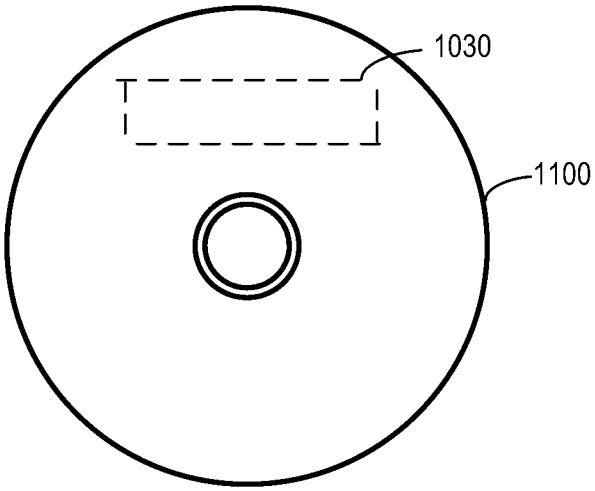

1030
1100

FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/095578

A. CLASSIFICATION OF SUBJECT MATTER

G06T 3/00(2006.01)i; G06T 7/00(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06T, G06K, G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, ENTXT, ENTXTC, DWPI, CNKI: IMAGE, DISTANCE, ABNORMAL, DISTANCE ESTIMATION, VEHICLE, TRACK, TRAIN, PERSPECTIVE PROJECT TRANSFORMATION, THREE DIMENSIONAL, TRAFFIC, MONITOR, ALARM, WARN, CORNER DETECTION

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102254318 A (UNIV SHANGHAI JIAOTONG) 23 November 2011 (2011-11-23) description, paragraphs [0045]-[0073], and figures 1-6	1-6, 11-17, 22,23
A	CN 102254318 A (UNIV SHANGHAI JIAOTONG) 23 November 2011 (2011-11-23) description, paragraphs [0045]-[0073], and figures 1-6	7-10, 18-21
A	CN 113836964 A (BEIJING TUSIMPLE FUTURE TECHNOLOGY CO) 24 December 2021 (2021-12-24) the whole document	1-23
A	CN 114663507 A (SHENZHEN SENSETIME TECHNOLOGY CO LTD) 24 June 2022 (2022-06-24) the whole document	1-23
A	WO 2021110497 A1 (VALEO SCHALTER & SENSOREN GMBH) 10 June 2021 (2021-06-10) the whole document	1-23

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 October 2023	**02 January 2024**

Name and mailing address of the ISA/CN	Authorized officer
CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	**HU,Yan** Telephone No. **(+86) 62089101**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/095578

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 102254318 A	23 November 2011	CN 102254318 B	09 January 2013
CN 113836964 A	24 December 2021	None	
CN 114663507 A	24 June 2022	None	
WO 2021110497 A1	10 June 2021	DE 102019132996 A1	10 June 2021