



(51) International Patent Classification:

G06T 7/593 (2017.01) H04N 13/128 (2018.01)
G06V 10/82 (2022.01)

(21) International Application Number:

PCT/EP2023/079921

(22) International Filing Date:

26 October 2023 (26.10.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

202241064626 11 November 2022 (11.11.2022) IN
2305580.9 17 April 2023 (17.04.2023) GB

(71) Applicant: **CONTINENTAL AUTONOMOUS MOBILITY GERMANY GMBH** [DE/DE]; Ringlerstraße 17, 85057 Ingolstadt (DE).

(72) Inventors: **KANNAN, Srividhya**; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG).

DHARMALINGAM, Ramalingam; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG). **HEGDE, Sneha K**; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG). **TANKSALE, Tejas**; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG). **VASHISHT, Abhishek**; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG). **HEINRICH, Stefan**; c/o Continental Automotive Singapore Pte. Ltd., Intellectual Property, 80 Boon Keng Road, Continental Building, Singapore 339780 (SG).

(74) Agent: **CONTINENTAL CORPORATION**; c/o Continental Automotive Technologies GmbH, Intellectual Property, Guerickestr. 7, 60488 Frankfurt am Main (DE).

(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,

(54) Title: IMAGE PROCESSING METHOD AND METHOD FOR PREDICTING COLLISIONS

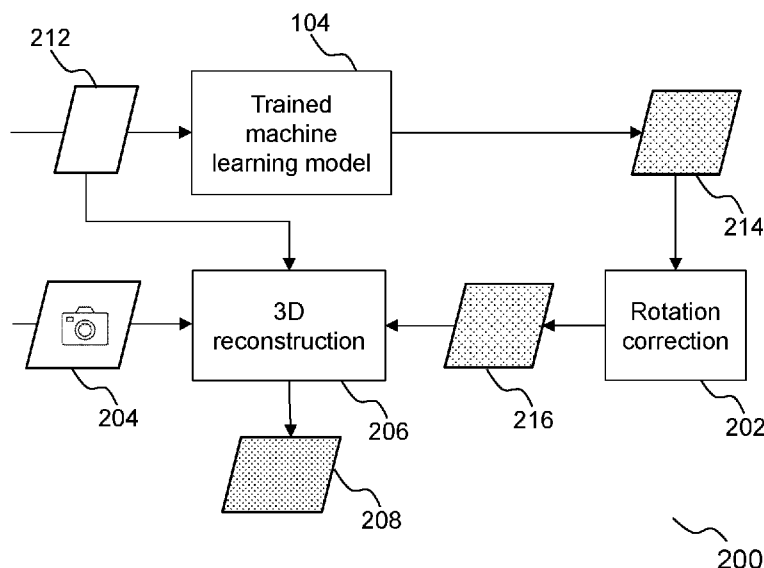


FIG. 2

(57) Abstract: According to various embodiments, a computer-implemented image processing method may include inputting a stereo image set to a machine learning model. The stereo image set may include a first image and a second image. The image processing method may further include generating a rotated disparity map of the stereo image set, using the machine learning model. The image processing method may further include, for each pixel in the first image - identifying a corresponding pixel in the second image based on the rotated disparity map, determining a respective rotation angle between the pixel in the first image and the corresponding pixel in the second image relative to a common image center, and determining a respective unrotated disparity based on the respective rotation angle. The image processing method may further include generating a unrotated disparity map based on the respective unrotated disparities of each pixel in the first image.



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, 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))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

IMAGE PROCESSING METHOD AND METHOD FOR PREDICTING COLLISIONS

TECHNICAL FIELD

[0001] Various embodiments relate to methods and devices for processing stereo images, methods and devices for predicting collisions and methods for training a machine learning model.

BACKGROUND

[0002] Stereovision techniques typically use two cameras to look at the same object. The two cameras may be separated by a baseline distance. The baseline distance is assumed to be known accurately. The two cameras may simultaneously capture two images, also referred to as stereo images. The stereo images may be analyzed to identify the differences between the two images. The differences between the two images may be referred to as disparity. The disparity between the two images may be used to determine depth of a point in the images. The depth information may be used to project the point in a three-dimensional model. Such three-dimensional (3D) models may be useful for facilitating various driver assistance or autonomous driving functions. For example, the three-dimensional model may provide information on positions of various objects near to a vehicle, thereby aiding navigation or obstacle avoidance.

[0003] For automobile applications, the stereo cameras may be installed onto a vehicle to capture images of the surroundings of the vehicle. The stereo cameras need to be precisely calibrated in order for the disparity acquired from the images to be accurate. However, regular movements of the vehicle, for example, over a pothole on the road, or over a road hump, may result in vibrations that displace the stereo cameras, thereby decalibrating the stereo cameras. When this happens, the stereo images cannot be relied on to generate accurate depth information that is needed for the three-dimensional model. Consequently, the three-dimensional model would not be suitable for use as a means to detect objects present in the vehicle's surroundings, or to prevent collisions with the objects.

SUMMARY

[0004] According to various embodiments, there is provided a computer-implemented image processing method. The image processing method may include inputting an image set to a machine learning model. The image set may include a first image and a second image. The image processing method may further include generating a rotated disparity map of the image set, using the trained machine learning model. The image processing method may further include, for each pixel in the first image – determining, based on the rotated disparity map, a rotation angle between the pixel in the first image and its corresponding position in the second image, relative to a common image center, and determining a respective unrotated disparity based on the respective rotation angle. The image processing method may further include generating an unrotated disparity map based on the respective unrotated disparity of each pixel in the first image.

[0005] According to various embodiments, there is provided an image processing device that includes a processor. The processor may be configured to perform the abovementioned image processing method.

[0006] According to various embodiments, there is provided a computer-implemented method for predicting collisions. The method may include inputting a sequence of image sets to a motion detection model. Each image set of the sequence may include a first image and a second image. The method may further include determining by the motion detection model, for each image set, a respective depth map based on the first image and the second image of the stereo image set, resulting in a sequence of depth maps. The method may further include determining by the motion detection model, an optical flow based on at least one of, the first images from the sequence of image sets and the second images of the sequence of stereo image sets. The method may further include determining by the motion detection model, motion of an object in the image set based on the optical flow. The method may further include determining time-to-collision with the object based on the sequence of depth maps and the determined motion of the object.

[0007] According to various embodiments, there is provided a device for predicting collisions. The device may include a processor configured to perform the abovementioned method for predicting collisions.

[0008] Additional features for advantageous embodiments are provided in the dependent claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] In the drawings, like reference characters generally refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead generally being placed upon illustrating the principles of the invention. In the following description, various embodiments are described with reference to the following drawings, in which:

[0010] FIG. 1 shows a block diagram of a computer-implemented method of training a machine learning model to determine disparity of images, according to various embodiments.

[0011] FIG. 2 shows a block diagram of a computer-implemented method of generating a 3D model using stereo images, according to various embodiments.

[0012] FIG. 3 shows an example of an image set according to various embodiments.

[0013] FIG. 4 shows an example of how a position of a feature point may differ in the left image and the right image.

[0014] FIG. 5A shows an example of an image set according to various embodiments.

[0015] FIG. 5B shows disparity images generated by a prior art machine learning model and the trained machine learning model according to various embodiments.

[0016] FIG. 6 shows an image processing method according to various embodiments.

[0017] FIG. 7A shows a simplified block diagram of an image processing device according to various embodiments.

[0018] FIG. 7B illustrates an operation of the image processing device according to various embodiments.

[0019] FIG. 8A shows a simplified functional block diagram of a device for predicting collisions according to various embodiments.

[0020] FIG. 8B shows a simplified hardware block diagram of the device according to various embodiments.

[0021] FIG. 9 shows a schematic diagram of the device performing a method for predicting collisions, according to various embodiments.

[0022] FIG. 10 shows a schematic diagram of a neural network according to various embodiments.

[0023] FIG. 11 shows a schematic diagram of the neural network according to various embodiments, receiving different inputs from those in FIG. 10.

[0024] FIG. 12 shows a flow diagram of a computer-implemented method for predicting collisions, according to various embodiments.

DESCRIPTION

[0025] Embodiments described below in context of the devices are analogously valid for the respective methods, and vice versa. Furthermore, it will be understood that the embodiments described below may be combined, for example, a part of one embodiment may be combined with a part of another embodiment.

[0026] It will be understood that any property described herein for a specific device may also hold for any device described herein. It will be understood that any property described herein for a specific method may also hold for any method described herein. Furthermore, it will be understood that for any device or method described herein, not necessarily all the components or steps described must be enclosed in the device or method, but only some (but not all) components or steps may be enclosed.

[0027] The term “coupled” (or “connected”) herein may be understood as electrically coupled or as mechanically coupled, for example attached or fixed, or just in contact without any fixation, and it will be understood that both direct coupling or indirect coupling (in other words: coupling without direct contact) may be provided.

[0028] In this context, the device as described in this description may include a memory which is for example used in the processing carried out in the device. A memory used in the embodiments may be a volatile memory, for example a DRAM (Dynamic Random Access Memory) or a non-volatile memory, for example a PROM (Programmable Read Only Memory), an EPROM (Erasable PROM), EEPROM (Electrically Erasable PROM), or a flash memory, e.g., a floating gate memory, a charge trapping memory, an MRAM (Magnetoresistive Random Access Memory) or a PCRAM (Phase Change Random Access Memory).

[0029] In order that the invention may be readily understood and put into practical effect, various embodiments will now be described by way of examples and not limitations, and with reference to the figures.

[0030] FIG. 1 shows a block diagram of a computer-implemented method 100 of training a machine learning model 102 to determine disparity of images, according to various embodiments. The method of training the machine learning model 102 may include providing training image data 112 to the machine learning model 102. The training image data 112 may include left stereo images and right stereo images from uncalibrated stereo cameras. In other words, the training image data 112 may include uncalibrated stereo images. Ground truth disparity data 114 that corresponds to each set pair of left and right stereo images of the training image data 112 may also be provided to the machine learning model 102 as a training signal. The machine learning model 102 may be configured to extract features from the training image data 112, and may learn a pattern between the extracted features and the ground truth disparity data 114. As a result of training using the training image data 112 and the ground truth disparity data 114, the machine learning model 102 learns to compute a disparity map for each pair of left and right stereo image pair. The resulting machine learning model is referred herein as the trained machine learning model 104. The trained machine learning model 104 may include a neural network 1000 that will be described further with respect to FIG. 10.

[0031] In a general stereo camera set up, a feature point of the left stereo image and the corresponding feature point of the right stereo image are aligned on the same x -axis, also referred to as horizontal axis. Due to vibration or mechanical set up, one of the stereo cameras may be misaligned with the other stereo camera with at least one of the following factors: roll, i.e. rotation angle (α), pitch, yaw, and translation (x_t, y_t) where x_t refers to movement of an image pixel along the horizontal axis and y_t refer to movement of the image pixel along the vertical axis. The trained machine learning model 104 may be configured to determine rotated disparities for the above-mentioned misalignments, as the training image data 112 includes uncalibrated stereo images. The method 100 may further include rotating stereo images to artificially uncalibrate the training image data 112.

[0032] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the training image data 112 may include images obtained from a plurality of cameras that may not be stereo cameras. These images may capture a similar scene from different angles, and hence, may similarly be used to determine depth of objects in the scene, like stereo images.

[0033] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the cameras or stereo cameras

that capture the training image data 112 may be coupled to a vehicle. The training image data 112 may capture scenes around the vehicle.

[0034] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the method 100 may further include classifying the disparity range into near range and far range while training the machine learning model 102. For example, the ground truth disparity data 114 may be split into a near range ground truth disparity data and a far range ground truth disparity data. The machine learning model 102 may be trained to compute near range ground truth disparity using the near range ground truth disparity data as the training signal. The machine learning model 102 may be further trained to compute far range ground truth disparity using the far range ground truth disparity data as the training signal, in a separate process from the training using near range ground truth disparity. By training the machine learning model 102 for near range and far range disparities separately, the accuracy of the trained machine learning model 104 may be improved. The time taken to train the machine learning model 102 may also be reduced.

[0035] FIG. 2 shows a block diagram of a computer-implemented method 200 of generating a 3D model using stereo images, according to various embodiments. The method 200 may include providing image data to the trained machine learning model 104. The image data may include an image set 212. The image set 212 may include a left stereo image and a right stereo image. The trained machine learning model 104 may extract feature information from the left and right stereo images, and may generate a rotated disparity map 214 based on the extracted feature information. A rotation correction module 202 may receive the rotated disparity map 214 and correct the rotated disparity map 214 for the relative rotation between the images in the image set 212. The rotation correction module 202 may generate a unrotated disparity map 216 based on the rotated disparity map 214. The rotation correction module 202 may provide the unrotated disparity map 216 to a 3D reconstruction module 206. The 3D reconstruction module 206 may generate a 3D point cloud 208 based on the unrotated disparity map 216, camera parameters 204 and the image set 212. The rotated disparity map 214 may include the actual disparities of every pixel between the uncalibrated left and right stereo images. The unrotated disparity map 214 may include corrected disparities of every pixel between the uncalibrated left and right stereo images, as if the misaligned stereo image was already corrected back to its calibrated position.

[0036] The method 200 may be able to generate 3D point clouds with depth measurements that are more accurate, and at a lower cost, as compared to LiDAR sensors. The method 200 may

also be able to determine the depth in scenes captured on the images regardless of dynamic decalibrations of the cameras, including rotation, horizontal shift and vertical shift. The method 200 may also be able to generate the 3D point clouds even if the image set 212 includes distorted images and modified intrinsics, and if the camera parameters are inaccurate, for example, incorrect focal lens information.

[0037] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image set 212 may include images obtained from a plurality of cameras that may not be stereo cameras. These images may capture a similar scene from different angles, and hence, may similarly be used to determine depth of objects in the scene, like stereo images.

[0038] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the cameras or stereo cameras that capture the image set 212 may be coupled to a vehicle. The image set 212 may capture scenes around the vehicle.

[0039] FIG. 3 shows an example of an image set 212 according to various embodiments. The image set 212 may include a left image 302 and a right image 304. The image set 212 may be uncalibrated. The right image 304 is rotated relative to the left image 302. This can happen, when at least one of the stereo cameras is not installed properly, or is displaced from its original position. For example, when the vehicle drives over a road hump or makes a sharp movement, the stereo camera may experience vibration that results in a slight displacement. The rotation correction module 202 may determine the unrotated disparity map 216 based on the rotated disparity map 214, such that the unrotated disparity map 216 provides information on the displacement between each pixel in the left image 302 and its corresponding right image 304 as if the left image 302 and the right image 304 are aligned on a horizontal axis. The process of determining the unrotated disparity map 216 is described with respect to FIG. 4.

[0040] FIG. 4 shows an example of how a position of a feature point may differ in the left image 302 and the right image 304. In this example, an image centre 402 may be considered to be (0,0). A feature point 404 of the left image 302 is denoted as (x_2, y_2) . The corresponding position 406 of the feature point 404 in the right image 304 as rotated, is denoted as (x'_2, y'_2) . A line connecting the corresponding position 406 and the image center 402 may define a first hypotenuse 420, indicated herein as " h_1 ", of a right angle triangle 410. The corresponding position 406 of the feature point 404 in the right image 304 may be determined based on the rotated disparity map 214.

[0041] The rotated disparity map may include rotated disparity values (dx_2, dy_2) for the feature point 404. The corresponding position 406, and the first hypotenuse 420, may be determined based on the feature point 404 and the rotated disparity values according to the following equations (1) to (3).

$$x'_2 = x_2 + dx_2 \quad \text{Equation (1)}$$

$$y'_2 = y_2 + dy_2 \quad \text{Equation (2)}$$

$$h_1 = \sqrt{(x'_2)^2 + (y'_2)^2} \quad \text{Equation (3)}$$

[0042] The unrotated disparity values of the feature point 404 may be expressed as (dx_2, dy_2). The unrotated corresponding position 408 of the feature point 404 in the right image 304 when the right image 304 is adjusted to have zero rotation angle relative to the left image 302, is denoted as (x_{2r}, y_{2r}).

[0043] The corresponding position 406, the unrotated corresponding position 408 and the image centre 402 may define vertices of an isosceles triangle 422. The distance between the corresponding position 406 and the image center 402 may be at least substantially equal to the distance between the unrotated corresponding position 408 and the image centre 402, in other words, equal to the first hypotenuse 420. The vertex angle 426 of the isosceles triangle 422 is denoted as a . The vertex angle 426 may also be referred herein as rotation angle 426.

[0044] A line connecting the corresponding position 406 and the unrotated corresponding position 408, may define a second hypotenuse 424, denoted herein as " h_2 ", of another right angle triangle 412. The second hypotenuse 424 may be the base of the isosceles triangle 422. The second hypotenuse 424 may be determined according to equation (4).

$$h_2 = 2 \times h_1 \times \sin \frac{a}{2} \quad \text{Equation (4)}$$

[0045] The base 428 of the other right angle triangle 412 is denoted as d . The length of the base 428 may be determined according to the equation (5).

$$d = \sqrt{(h_2^2 - dy_2^2)} \quad \text{Equation (5)}$$

[0046] The unrotated disparity value d_2 may be determined based on the base 428 and the rotated disparity value dx_2 . The unrotated disparity value may be determined according to the following equation (6).

$$d_2 = d + dx_2$$

Equation
(6)

[0047] The rotation correction module 202 may determine the unrotated disparity map 216 based on the abovementioned computations. The rotation correction module 202 may determine the corresponding position 406 based on the rotated disparity values in the rotated disparity map 214. The rotation correction module 202 may determine the first hypotenuse 420 based on the corresponding position 406. The rotation correction module 202 may determine the vertex angle 426. Determining the vertex angle 426 may include determining direction of epipolar lines of the image set 212, in the image plane without using 3D space. The vertex angle 426 may then be obtained by projecting the rotated disparity measurements towards the epipolar line directions. The rotation correction module 202 may determine the second hypotenuse 424 based on the vertex angle 426 and the first hypotenuse 420. The rotation correction module 202 may determine the base 428 based on the second hypotenuse 424 and the rotated disparity values. The rotation correction module 202 may determine the unrotated disparity value based on the base 428 and the rotated disparity values.

[0048] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the rotation correction module 202 may be configured to determine direction of epipolar lines of the image set 212, in the image plane without using 3D space. The rotation correction module 202 may determine the epipolar line directions without prior knowledge about the intrinsic and extrinsic camera parameters. The rotation correction module 202 may determine the epipolar line directions, by processing the images in the image set 212, region by region. In other words, the images may be segmented into a plurality of regions, and the epipolar line directions may be determined in each region of the plurality of regions.

[0049] The rotation correction module 202 may be configured to check for infinity-negative disparities in the vertical disparity. The real distance represented in the images may be computed by the projection of the disparities through the epipolar line direction. The real distances may be determined region by region, in other words, determined in each region of the plurality of regions.

[0050] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the rotation correction module 202 be further configured to determine the translation (x_t, y_t) and further configured to correct the right image with respect to the left image, based on the determined translation. The

horizontal translation x_t may be computed using tracking and filtering of the input images over time. The vertical translation y_t may be determined based on the pixel information in the centre of y-disparity in a y-disparity array.

[0051] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the 3D reconstruction module 206 may be configured to generate an unscaled 3D point cloud based on the unrotated disparity map 216. The trained machine learning model 104 may receive a sequence of image sets 212 over time, and thereby generating a sequence of rotated disparity maps 214. Accordingly, the rotation correction module 202 may generate a sequence of unrotated disparity maps. The 3D reconstruction module 206 may correct the 3D point cloud for camera factors such as scale and yaw angle, based on comparing at least two consecutive 3D point clouds. Points in the 3D point cloud should move at least substantially at the same velocity, as the points move relative to the vehicle that the cameras are coupled to. In other words, the points in the 3D point cloud may move at a velocity that is at least substantially equal to the longitudinal speed of the vehicle. The 3D reconstruction module 206 may correct the 3D point cloud for camera factors based on velocity deviations of points in the 3D point cloud.

[0052] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the method 200 may further include correcting for negative disparities in the image set 212. The correction for negative disparities may be performed before performing 3D reconstruction. The negative disparities may be caused by yaw angle errors between the cameras that captured the image set 212. A change in the yaw angle of at least one camera may result in a horizontal shift. In stereo images, the horizontal disparity decreases as distance, i.e. depth, increases. As such, the horizontal disparity of objects in infinity, also referred herein as infinity objects, should be at least substantially zero. However, the infinity objects may display negative horizontal disparities instead of zero disparity when there is a negative yaw angle between the cameras. The method 200 may further include applying a low pass filter to the unrotated disparity map 216, to obtain the maximum negative disparity values, i.e. the negative disparity values with the largest amplitude, in the unrotated disparity map 216. The method 200 may include using the maximum negative disparity value to correct the disparities in the entire image region. Correction of the disparities may be performed region by region, where each image may be divided into a plurality of regions.

[0053] FIG. 5A shows an example of an image set 212 according to various embodiments. In this example, the image set 212 includes a first image 502 and a second image 504. The first image 502 was taken by a first camera while the second image 504 was taken by a second camera. Both the first camera and the second camera captured images of the same scene, from slightly offset positions. In other words, the second camera is positioned at a calibrated distance away from the first camera. In a scenario when the second camera is decalibrated in terms of rotation, the second image 504 is rotated relative to the first image 502.

[0054] FIG. 5B shows disparity images generated by a prior art machine learning model and the trained machine learning model 104 according to various embodiments. The disparity images include a first disparity image 510 and a second disparity image 520, which were both generated by machine learning models based on the example image set 212 of FIG. 5A. The first disparity image 510 was generated by a prior art machine learning model. The prior art machine learning model failed to handle the decalibration of the second image 504, and as such, objects in the image set 212 are not visible in the first disparity image 510. In contrast, the second disparity image 520 clearly shows the objects in the image set 212, which indicates that the trained machine learning model 104 was able to correctly match features between the first image 502 and the second image 504 in spite of the decalibration of the second image 504.

[0055] FIG. 6 shows an image processing method 600 according to various embodiments. The image processing method 600 may include processes 602, 604, 606, 608 and 610. The process 602 may include inputting an image set 212 to a trained machine learning model 104. The image set 212 may include a first image and a second image. The process 604 may include generating a rotated disparity map 214 of the image set 212, using the trained machine learning model 104. The process 606 may include, for each pixel in the first image, determining based on the rotated disparity map 214, a rotation angle between the pixel in the first image and its corresponding position in the second image, relative to a common image centre. The process 608 may include, for each pixel in the first image, determining a respective unrotated disparity based on the respective rotation angle. The process 610 may include generating unrotated disparity map 216 based on the respective unrotated disparities of each pixel in the first image. The image processing method 600 may determine an unrotated disparity map that may be used to accurately determine depth of points in the image set 212, for example, to construct an accurate 3D point cloud, without having to calibrate the cameras that capture the image set 212. These cameras may be mounted on a vehicle, to capture the surroundings of the vehicle. Cameras mounted on the vehicle may shift out of their initial calibrated positions due to

vibrations or movements of the vehicle. Using the image processing method 600, the depth of points in the image set 212 may be determined accurately without being affected by the decalibration of the cameras.

[0056] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image set 212 may include stereo images. The first image may be a left stereo image, while the second image may be a right stereo image. A stereo image set may capture a similar scene from slightly different positions, such that the images within the stereo image set may be compared to obtain depth information on each point in the images. This depth information may be useful for providing situational awareness to a vehicle, such as indicating the vehicle's distances to various objects in its surroundings.

[0057] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the trained machine learning model 104 may be configured to determine both horizontal disparities and vertical disparities of the image set 212, such that the rotated disparity map 214 may include both horizontal disparities and vertical disparity values of the image set. By determining both horizontal disparities and vertical disparities of the image set 212, the trained machine learning model 104 may be capable of handling decalibration of the cameras in both vertical and horizontal directions. As such, the image processing method 600 may result in an accurate disparity map even if at least one of the cameras is shifted out of place in both directions, or is rotated.

[0058] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the common image centre may be an arbitrary point in the image and may be close to the image centre of at least one of the first image and the second image.

[0059] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, a centre of the first image may be used as the common image centre. The first image may be used as a reference image, and the translation of the second image relative to the first image may be estimated.

[0060] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing method 600 may further include segmenting the first image into a plurality of first regions, and segmenting the second image into a plurality of second regions corresponding to the plurality of first regions. The image processing method 600 may further include determining an epipolar

line and its direction based on a pixel in the first image and its corresponding position in the second image, for each first region of the plurality of first regions.

[0061] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing method 600 may further include for each pixel of the first image, determining a respective horizontal disparity based on projection of the respective unrotated disparity through the direction of the epipolar line. The horizontal disparity in the image set 212, for example, caused by misalignment of at least one camera, may be thereby corrected.

[0062] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing method 600 may further include generating a 3D point cloud 208 based on the image set 212 and further based on the unrotated disparity map 216. The 3D point cloud may provide detailed information about the surroundings of a vehicle, and may serve to aid the vehicle in various functions such as navigation, localization and avoidance of obstacles.

[0063] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing method 600 may further include inputting a further image set to the trained machine learning model 104. The further image set may include a further first image and a further second image captured at a consecutive time frame from the first image and the second image of the image set 212. The image processing method 600 may further include generating a further rotated disparity map of the further image set, using the trained machine learning model 104. The image processing method 600 may further include, for each pixel in the further first image, identifying a corresponding pixel in the further second image based on the further rotated disparity map, determining a respective rotation angle between the pixel in the further first image and the corresponding pixel in the further second image, relative to a further common image centre, and determining a respective unrotated disparity based on the respective rotation angle. The image processing method 600 may further include generating a further unrotated disparity map based on the respective unrotated disparities of each pixel in the further first image. The image processing method 600 may further include generating a further 3D point cloud based on the further image set and further based on the further unrotated disparity map, determining a distance that each point in the 3D point cloud moves from the 3D point cloud 208 to the further 3D point cloud, and correcting at least one of the 3D point cloud 208 and the further 3D point cloud, based on the determined distances. The 3D reconstruction module 206

may perform the above-described correction of the 3D point cloud 208 or the further 3D point cloud. Points in the 3D point cloud should move at least substantially at the same velocity, as the points move relative to the vehicle that the cameras are coupled to. The 3D reconstruction module 206 may correct the 3D point cloud for camera factors such as scale and yaw angle, based on velocity deviations of points in the 3D point cloud.

[0064] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the trained machine learning model 104 may be trained using a training dataset 112 that includes a plurality of image pairs generated by rotating a pair of calibrated images to a corresponding plurality of different angles, and using a ground truth disparity map 114 generated based on the pair of calibrated images as a training signal. As the training dataset 112 is generated using calibrated images, a precise ground truth disparity map 114 may be obtained.

[0065] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the trained machine learning model 104 may be trained to determine near range disparities, and further separately trained to determine far range disparities. Doing so may optimize the training process, reducing the training time required and improving the accuracy of the trained machine learning model 104 in determining the disparities.

[0066] FIG. 7A shows a simplified block diagram of an image processing device 700 according to various embodiments. The image processing device 700 may include a processor 702. The processor 702 may be configured to perform the image processing method 600.

[0067] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing device 700 may be any one of a server, a computer, a vehicle or a robot. The image processing device 700 may be an autonomous vehicle, such as a self-driving car, or a drone.

[0068] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the image processing device 700 may further include at least one camera 704. The camera 704 may be configured to capture the image set 212. The image processing device 700 may further include an engine 706 configured to drive the image processing device 700. The image processing device 700 may further include a steering module 708 configured to steer the image processing device 700.

[0069] FIG. 7B illustrates an operation of the image processing device 700 according to various embodiments. The input 720 to the image processing device 700 may include a

sequence of image sets. The sequence of image sets may include consecutively captured images. For example, the sequence of image sets may include an image set 720a captured at $t-1$, and another image set 720b captured at t , where t denotes time as a variable. In other words, the image set 720b may be a subsequent frame to the image set 720a. Each image set may include a plurality of images, each captured at a respective position. These positions may be offset from one another, such that a combination of the plurality of images may provide depth information of the objects shown in the images. For example, each of the image set 720a and the image set 720b may include a pair of stereo images. The image set 720a may include a left image 722a and a right image 724a. The image set 720b may include a left image 722b and a right image 724b.

[0070] The sequence of image sets 720a, 720b may be provided to the trained machine learning model 104. The trained machine learning model 104 may be trained to compute a respective rotated disparity map 214 for each image set. The trained machine learning model 104 may be trained to identify features in images and further configured to determine spatial offset, also referred herein as disparity data, of the features between the images. The spatial offset between images of the same image set may provide depth information of the features. A rotation correction module 202 may compute a respective unrotated disparity map 216 for each rotated disparity map 214, for example, like described with respect to FIG. 4.

[0071] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the device 700 may further include a point cloud generator 726. The point cloud generator 726 may include the 3D reconstruction module 206. The point cloud generator 726 may be configured to generate a 3D point cloud 208 based on the input 720 and the unrotated disparity map 216. The generated 3D point cloud 208 may be a 3D reconstruction of the environment that a vehicle is travelling in. As the 3D point cloud 250 may be generated in real-time, the device 100 may provide the vehicle with environmental data that is not previously available, for example, previously unmapped terrain. Further, the 3D point cloud 208 may indicate to the vehicle, the presence of dynamic objects such as other traffic participants. The 3D point cloud 208 may be generated further based on camera intrinsic parameters such as focal length along x, y camera principal point offset and axis skew. The point cloud generator 726 may also refine the 3D point cloud 208 to remove outliers and incorrect predictions, so that the 3D point cloud 208 may serve as an accurate dense 3D map. The point cloud generator 726 may remove the outliers and incorrect predictions based on probability of those data points.

[0072] Various aspects described with respect to the method 200 and the image processing method 600 may be applicable to the image processing device 700.

[0073] Collision prediction plays an important role in improving traffic safety. The majority of past research focused on determining time-to-collision (TTC) during the day when oncoming vehicles are visible. However, traffic accidents are more frequent at night since there is less visual information about vehicles and complex lighting conditions. At night, the visual feature information is almost reduced to the headlights and taillights of the vehicles. Traditional collision prediction systems rely on the vehicle's width and motion information, which is difficult to assess under poor lighting. Existing deep learning models trained for vehicle light recognition at night generally fail to generate reliable TTC values on pitch dark images. In advanced driver assistance systems, several algorithms rely solely on these features for vehicle detection. TTC is the most commonly used safety metric, and it indicates the time it takes for a vehicle to collide with another vehicle.

[0074] According to various embodiments, a computer-implemented method 1200 for predicting collisions is provided. The method 1200 may estimate TTC based on stereo disparity information and optical flow. The stereo disparity information may be obtained using the image processing method 600. The method 1200 is also described with respect to FIG. 12.

[0075] The method 1200 may include receiving a sequence of image sets. The sequence of image sets in the input 720 may be captured by sensors mounted on a vehicle. Examples of suitable sensors include stereo vision camera, thermal stereo camera, or dense LiDAR. LiDAR sensors may provide dense point cloud of the surrounding environment as well as the distance to each detected object in the scene. Each image set may include a plurality of images, and each image of the plurality of images may be captured from a different position on the vehicle, such that the plurality of images of each image set may have an offset in at least one axis, from one another. The plurality of images may be respectively captured by a corresponding plurality of sensors. The plurality of images may include a first image and a second image.

[0076] In an example, the sensor may be a pair of stereo cameras. As such, the image set includes a pair of stereo images. The two cameras may be separated by a baseline, the distance for which is assumed to be known accurately. The cameras may simultaneously capture two consecutive images. The images may be analyzed to identify differences between the images, and to identify the corresponding pixel-positions in both images, in a stereo matching process. The disparity between corresponding pixel in both images may be used to estimate depth. In

an example, the pair of stereo cameras may be mounted on the side mirror of a vehicle. The stereo camera may be FSC231 stereo camera with 8.3M Pixel resolution and 30° field of view.

[0077] The method 1200 may include determining the disparity between the images in the image set, using a motion detection module that includes a machine learning model. The machine learning model may be, for example, the trained machine learning model 104. The machine learning model may include a neural network, such as graph neural network, transformers, or recurrent neural network. The machine learning model may include the neural network 1000 described with respect to FIG. 10. Using the estimated disparity, the distance and TTC for each given pixel or object in the scene captured by the images may be computed. The distance may be measured in meters, while the TTC may be measured in seconds. The features extracted for performing the stereo matching may be used to estimate TTC for each pixel in the image using deep neural network.

[0078] The method 1200 may include recognizing moving objects, by identifying moving pixels in the sequence of image sets. In other words, recognizing moving objects may involve finding corresponding pixels in the sequence of image sets, over time. Recognition of the moving objection may be accomplished by passing two consecutive stereo image pairs through an optical flow estimation algorithm.

[0079] The method 1200 may include combining the resulting flow information with the current stereo frame to segment moving objects in a scene, using a neural network. A point cloud generator may generate a 3D point cloud of the surrounding of the vehicle, based on the estimated depth, i.e. distance of each pixel.

[0080] According to various embodiments, the method 1200 may include providing stereo camera data to a motion detection module. The motion detection module may include a machine learning model, also referred herein as motion detection network. The motion detection network may include the trained machine learning model 104. The motion detection network may include, for example, a Convolutional Neural Network (CNN). The stereo camera data may include a first stereo image pair and a second stereo image pair. The first stereo image pair and the second stereo image pair may be captured successively. The motion detection network may be trained using stereo camera data, ground truth disparity and ground truth TTC values. The stereo camera data may be images captured at night, so that the motion detection network may learn to compute the disparity values and TTC values for night images. The motion detection network may also be trained with images associated with other environmental conditions such as day light, rain or snow. The motion detection network may be trained with

uncalibrated stereo images, and may generate a rotated disparity map of the stereo camera data. The motion detection module may include a rotation correction module 202 that converts the rotated disparity map to unrotated disparity map. The motion detection network may perform stereo matching, by extracting feature information of left and right stereo images, and mapping each image to a dense disparity map. The motion detection network may be trained to determine the rotated disparity, the optical flow, and the TTC image output of two consecutive stereo image pairs. In order to segment moving objects in a scene, the flow information obtained may be combined with a current stereo frame to train the motion detection network to predict the mask of moving objects. The motion detection module may determine the relative depth of the moving objects based on the unrotated disparity map, and may further determine a TTC image based on the determined relative depth. The TTC image may include TTC values for each pixel in the image. The motion detection module may further perform trajectory planning based on the TTC estimation.

[0081] According to various embodiments, the disparity map, the input image and the flow information may be further used for estimating pose and performing simultaneous localization and mapping (SLAM) of the scene, for navigation of the vehicle.

[0082] FIG. 8A shows a simplified functional block diagram of a device 800 for predicting collisions according to various embodiments. The device 800 may be configured to receive an input 720 and may be configured to generate an output. The input 720 may include a sequence of image sets. The output may include predicted time-to-collision (TTC) 820 of a vehicle with another object or vehicle. The device 800 may be capable of determining the TTC 820 based on processing visual data, i.e. images. The device 800 may be configured to perform the method 1200.

[0083] The sequence of image sets in the input 720 may be captured by sensors mounted on a vehicle. Each image set may include a plurality of images, and each image of the plurality of images may be captured from a different position on the vehicle, such that the plurality of images of each image set may have an offset in at least one axis, from one another. The plurality of images may be respectively captured by a corresponding plurality of sensors. The plurality of images may include a first image and a second image. In some embodiments, the first image and the second image may be a pair of stereo images.

[0084] The device 800 may include a motion detection module 802. The motion detection module 802 may include a machine learning model, for example, the trained machine learning model 104. The motion detection module 802 may be configured to generate a respective depth

map 812 of each image set, based on the first image and the second image of the image set. Consequently, the motion detection module 802 may output a sequence of depth maps 812 based on the received sequence of image sets. Each depth map 812 may be an image or image channel that contains information relating to the distance of the surfaces of objects from a viewpoint. The viewpoint may be a vehicle, or more specifically, a sensor mounted on the vehicle.

[0085] The motion detection module 802 may also be configured to determine an optical flow 814 based on at least one of, the first images from the sequence of image sets and the second images of the sequence of image sets. The motion detection module 802 may determine motion of an object in the image set based on the optical flow 814, to generate motion data 816.

[0086] The machine learning model in the motion detection module 802 may be used to determine both disparity and motion. The machine learning model may generate the segmented mask of moving objects based on the detected motion.

[0087] The device 800 may further include a prediction module 804. The prediction module 804 may be configured to receive the sequence of depth maps 812 and the motion data 816 from the motion detection module 802. The prediction module 804 may be configured to generate the predicted TTC 820 based on the received sequence of depth maps 812 and the motion data 816.

[0088] According to an embodiment which may be combined with any of the above-described embodiment or with any below described further embodiment, the motion detection module 802 may include the device 700. The motion detection module 802 may determine the depth maps 812 by generating a respective unrotated disparity map 216 for each image set, for example, according to the image processing method 600.

[0089] According to an embodiment which may be combined with any of the above-described embodiment or with any below described further embodiment, the device 800 may further include a point cloud generator 726 that generates a 3D point cloud 208 based on at least one image set of the sequence of image sets. The device 800 may be further configured to detect an object in the 3D point cloud 208.

[0090] According to an embodiment which may be combined with any of the above-described embodiment or with any below described further embodiment, the point cloud generator 726 may generate a respective 3D point cloud based on each image set of the sequence of image sets, resulting in a sequence of 3D point clouds 208. The device 800 may compare distances

moved by each point in the sequence of 3D point clouds 208, and may correct at least one 3D point cloud of the sequence of 3D point clouds 208 based on the determined distances.

[0091] FIG. 8B shows a simplified hardware block diagram of the device 800 according to various embodiments. The device 800 may include at least one processor 830. The at least one processor 830 may be configured to carry out the functions of the machine learning model 802 and the prediction module 804.

[0092] According to various embodiments, the device 800 may be a driver assistance system.

[0093] According to various embodiments, the device 800 may be a vehicle.

[0094] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the device 800 may further include a plurality of sensors 832. The plurality of sensors 832 may be configured to generate the sequence of image sets 816. The plurality of sensors 832 may include, for example, stereo cameras, surround view cameras, infrared cameras, event cameras, or dense LiDAR. The device 800 may include at least one memory 834. The at least one memory 834 may store the machine learning model 802 and the prediction module 804. The at least one memory 834 may include a non-transitory computer-readable medium. The at least one processor 830, the plurality of sensors 832 and the at least one memory 834 may be coupled to one another, for example, mechanically or electrically, via the coupling line 840.

[0095] FIG. 9 shows a schematic diagram of the device 800 performing a method for predicting collisions, according to various embodiments. The motion detection module 802 of the device 800 may include a machine learning model that may generate a sequence of depth maps 812, optical flow 814 and motion data 816, based on a received sequence of image sets. The sequence of image sets may include at least a first image set 720a and a second image set 720b. Each of the image sets may include at least a first image and a second image, for example left image 722a and right image 724a of the first image set 720a, and the left image 722b and the right image 724b of the second image set 720b. The motion detection module 802 may also receive a moving object mask 914, which may be generated by another machine learning model.

[0096] The motion detection module 802 may output a segmented moving object image 910 based on the moving object mask 914 applied to the sequence of image sets. The device 800 may also include a tracking moving object module 906 configured to determine the motion data 816 based on the segmented moving object image 910.

[0097] The prediction module 804 of the device 800 may include a relative depth estimation module 902. The relative depth estimation module 902 may be configured to receive the sequence of depth maps 812, the optical flow 814 and the motion data 816. The relative depth estimation module 902 may determine relative depth of pixels in the images, based on the received inputs. The prediction module 804 may further include a TTC determination module 904. The TTC determination module 904 may determine the TTC of each pixel, based on the determined relative depth and the motion data 816.

[0098] FIG. 10 shows a schematic diagram of a neural network 1000 according to various embodiments. The neural network 1000 may be part of at least one of the trained machine learning model 104 and the motion detection module 802. The neural network 1000 may include a feature extraction network 1310 and a disparity computation network 1320. The feature extraction network 1310 may be configured to extract features from images in the sequence of image sets to generate feature maps. The disparity computation network 1320 may be configured to determine two-dimensional offsets (also referred herein as displacements or disparities) between the images based on the generated feature maps. The neural network 1000 may thereby determine both optical flow and depth information using a single common set of neural networks, as both the optical flow and depth information relate to two-dimensional offsets between images. As such, the neural network 1000 may be trained in a shorter time and with less resources, as compared to training two separate neural networks.

[0099] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the neural network 1000 may be trained via supervised training, using scene flow stereo images. Training the neural network 1000 may require, for example, 25,000 scene flow stereo images as the training data. The neural network 1000 may be fine-tuned using stereo images from the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset. As an example, about 400 stereo images from the KITTI dataset may be used.

[00100] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the feature extraction network 1310 may include a plurality of neural network branches. For example, the neural network branches may include a first branch 1350 and a second branch 1360. Each neural network branch may include a respective convolutional stack, and a pooling module connected to the convolutional stack. The convolutional stack may include, for example, the CNN 1312. The pooling module may include, for example, the SPP module 1314. The plurality of neural

network branches may share the same weights. By having the neural network branches share the same weights, the feature extraction network 1310 may be trained in a shorter time and with less resources, than training the neural network branches individually.

[00101] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the disparity computation network 1320 may include a 3D convolutional neural network (CNN) 1324 configured to generate three disparity maps based on the generated feature maps 1318. The feature extraction network 1310 may extract features at different levels. To aggregate the feature information along disparity dimension as well as spatial dimensions, the 3D CNN 1324 may be configured to perform cost volume regularization on the extracted features. The 3D CNN 11324 may include an encoder-decoder architecture including a plurality of 3D convolution and 3D deconvolution layers with intermediate supervision. The 3D CNN 324 may have a stacked hourglass architecture including three hourglasses, thereby producing three disparity outputs. The architecture of the 3D CNN 1324 may enable it to generate accurate disparity outputs and to predict the optical flow. The filter size in the 3D CNN 1324 may be 3*3.

[00102] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the neural network 1000 may be configured to determine the respective depth map 812 of each image set based on the determined 2D offsets between the plurality of images of the image set. The depth map 812 may provide information on distance between objects in the images from the vehicle. This information may improve the localization accuracy. Each pixel in the depth map may be determined based on the following equation:

$$Depth = \frac{baseline \times focal\ length}{disparity}$$

where baseline refers to the horizontal distance between the viewpoints that the images within the same image set were captured, focal length refers to distance between the lens and the image sensor, and disparity refers to the 2D offsets. The 2D offsets may be obtained from the unrotated disparity map 216.

[00103] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the neural network 1000 may be configured to determine the optical flow 814, based on the determined 2D offsets between at least one of, the first images of adjacent image sets of and the second images of

adjacent image sets. The optical flow 814 may provide information on the changes in position of objects in the images.

[00104] The neural network 1000 may be configured to perform a stereo matching operation. The stereo matching operation may include determining estimate a pixelwise displacement map between the input images. The input images may include a plurality of images of the same image set. For example, when the input contains images captured by a stereo camera, the input images may include a left image 722b and a right image 724b.

[00105] In general, stereo images may be rectified stereo images or unrectified stereo images. Rectified stereo images are stereo images where the displacement of each pixel is constrained to a horizontal line. The displacement map may be referred herein as disparity. To obtain rectified stereo images, the sensors or cameras used to capture the images need to be accurately calibrated.

[00106] Unrectified stereo images, on the other hand, may exhibit both vertical and horizontal disparity. Vertical disparity may be defined as the vertical displacement between corresponding pixels in the left and right images. Horizontal disparity may be defined as the horizontal displacement between corresponding pixels in the left and right images. It is challenging to obtain rectified stereo images from sensors mounted on a vehicle, as the sensors may shift or rotate in position over time, due to movement and vibrations of the vehicle. As such, the input images captured by the sensors mounted on the vehicle may be regarded as unrectified stereo images.

[00107] The neural network 1000 may be trained to be robust against rotation, shift, vibration and distortion of the input images. The neural network 1000 may be trained to handle both horizontal (x -disparity) and vertical (y -disparity) displacement. In other words, the neural network 1000 may be configured to determine both the horizontal and vertical disparity of the input images. This makes the neural network 1000 robust against vertical translation and rotation between cameras or sensors.

[00108] The neural network 1000 may have a dual-branch neural network architecture including a first branch 1350 and a second branch 1360, such that each branch may be configured to determine disparity in a respective axis. Each of the feature extract network 310 and the disparity computation network 1320 may include components of the first branch 350 and the second branch 360.

[00109] The feature extraction network 1310 may include a convolutional neural network (CNN) 1312, a spatial pyramid pooling (SPP) module 1314 and a convolution layer

1316, for each of the first branch 11350 and the second branch 1360. The CNN 1312 may extract feature information from the input images. The CNN 1312 may include three small convolution filters with kernel size (3×3) that are cascaded to construct a deeper network with the same receptive field. The CNN 1312 may include conv1_x, conv2_x, conv3_x, and conv4_x layers that form the basic residual blocks for learning the unitary feature extraction. For conv3_x and conv4_x, dilated convolution may be applied to further enlarge the receptive field. The output feature map size may be $(1/4 \times 1/4)$ of the input image size. The SPP module 1314 may be then applied to gather context information from the output feature map. The SPP module 1314 may learn the relationship between objects and its sub regions to incorporate hierarchical context information. The SPP module 1314 may include four fixed-size average pooling blocks of size 64×64 , 32×32 , 16×16 , and 8×8 . The convolution layer 1316 may be a 1×1 convolution layer for reducing feature dimension. The feature extraction network 1310 may up-sample the feature maps to the same size as the original feature map, using bilinear interpolation. The size of the original feature map may be $1/4$ of the input image size. The feature extraction network 1310 may concatenate the different levels of feature maps extracted by the various convolutional filters, as the left SPP feature map 1318 and the right SPP feature map 1319.

[00110] The disparity computation network 1320 may receive the left SPP feature map 1318 and the right SPP feature map 1319. The disparity computation network 1320 may concatenate the left and right SPP feature maps 1318, 1319 into separate cost volumes 1322 for x and y displacements respectively. Each cost volume may have 4 dimensions, namely, height $\times$ width $\times$ disparity $\times$ feature size. The disparity computation network 1320 may include a 3D-CNN 1322 in each branch. The 3D-CNN 1322 may include a stack hourglass (encoder-decoder) architecture that is configured to generate three disparity maps. The disparity computation network 1320 may further include an upsampling module 1326 and a regression module 1328, in each branch. The upsampling module 1326 may upsample the three disparity maps so that their resolution matches that of the input image size. The regression module 1328 may apply regression to the upsampled disparity maps, to calculate an output disparity map. The disparity computation network 1320 may calculate the probability of each disparity based on the predicted cost via SoftMax operation. The predicted disparity may be calculated as the sum of each disparity weighted by its probability. Next, smooth loss function may be applied between ground truth disparity and predicted disparity. The smooth loss function may measure how close the predictions are, to the ground truth disparity values. The smooth loss function

may be a combination of l1 and l2 loss. It is used in deep neural network because of its robustness and low sensitivity to outliers. The disparity computation network 1320 then outputs the horizontal displacement 1332 at the first branch 1350, and outputs the vertical displacement 1334 at the second branch 1360. The machine learning model 102 may then determine a depth map 112 based on the horizontal displacement 1332 and the vertical displacement 1334, using known stereo computation methods such as semi global matching.

[00111] FIG. 11 shows a schematic diagram of the neural network 1000 according to various embodiments, receiving different inputs from those in FIG. 10. The neural network 1000 may also be configured to perform optical flow computation. The optical flow computation may include predicting a pixelwise displacement field, such that for every pixel in a frame, the neural network 1000 may estimate its corresponding pixel in the next frame. For the optical flow computation operation, the input images used by the neural network 1000 may be two successive images taken from the same position. In this example, the input images are left image 722b captured at time = t and left image 722a captured at time = $t - 1$. The outputs of the neural network 1000 for the optical flow computation operation is the optical flow 814 that includes x -direction displacement and y -direction displacement.

[00112] The first branch 1350 may process the left image 722b, while the second branch 1360 may process the earlier left image 722a. Similar to the stereo-matching operation described with respect to FIG. 10, the optical flow computation operation may include extracting features using the CNN 1312 of the feature extraction network 1310. The SPP module 1314 may gather context information from the output feature map generated by the CNN 11312. The feature extraction network 1310 generate final SPP feature maps that are provided to the disparity computation network 1320. The difference from the stereo-matching operation, is that the SPP feature maps generated are an earlier SPP feature map (for time = $t - 1$) and a subsequent SPP feature map (for time = t). The disparity computation network 1320 may concatenate the SPP feature maps into separate cost volumes 1322 for time = $t - 1$ and time = t respectively. Each cost volume may have 4 dimensions, namely, height $\times$ width $\times$ disparity $\times$ feature size. The 3D-CNN 1322 of each branch may generate three disparity maps based on the respective cost volume 1322. The upsampling module 1326 may upsample the three disparity maps. The regression module 1328 may apply regression to the upsampled disparity maps, to calculate an output disparity map. The disparity computation network 1320 may calculate the probability of each disparity based on the predicted cost via SoftMax operation. The predicted disparity may be calculated as the sum of each disparity weighted by

its probability. Next, smooth loss function may be applied between ground truth disparity and predicted disparity. The disparity computation network 320 then outputs the x -direction displacement and the y -direction displacement, that are determined to take place between time $= t - 1$ to time $= t$.

[00113] According to various embodiments, suitable deep learning models for the machine learning model 102 may include, for example, PyramidStereoMatching and RAFTNet.

[00114] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, training of the neural network 1000 may, for example, be based on standard backpropagation based gradient descent. As an example how the neural network 1000 may be trained, a training dataset may be provided to the neural network 1000, and the following training processes may be carried out:

[00115] An example of a suitable training dataset for training the neural network 1000 may be the KITTI dataset.

[00116] Before training the neural network 1000, the weights may be randomly initialized to numbers between 0.01 and 0.1, while the biases may be randomly initialized to numbers between 0.1 and 0.9.

[00117] Subsequently, the first observations of the dataset may be loaded into the input layer of the neural network and the output value(s) is generated by forward-propagation of the input values of the input layers. Afterwards the following loss function may be used to calculate loss with the output value(s):

[00118] Mean Square Error (MSE): $MSE = \frac{1}{n} \sum (y - \hat{y})^2$, where n represents the number of neurons in the output layer and y represents the real output value and $\hat{y}$ represents the predicted output. In other words, $y - \hat{y}$ represents the difference between actual and predicted output.

[00119] The weights and biases may subsequently be updated by an AdamOptimizer with a learning rate of 0.001. Other parameters of the AdamOptimizer may be set to default values. For example:

beta_1 = 0.9

beta_2 = 0.999

eps = 1e-08

weight_decay = 0

[00120] The steps described above may be repeated with the next set of observations until all the observations are used for training. This may represent the first training epoch, and may be repeated until 10 epochs are done.

[00121] FIG. 12 shows a flow diagram of a computer-implemented method 1200 for predicting collisions, according to various embodiments. The method 1200 may include processes 1202, 1204, 1206, 1208 and 1210. The process 1202 may include inputting a sequence of image sets to a motion detection module, wherein each image set of the sequence comprises a first image and a second image. The sequence of image sets may include for example, the image sets 720a, 720b. The motion detection module may be for example, the motion detection module 802. The motion detection module may include the neural network 1000. The process 1204 may include determining by the motion detection module, for each image set, a respective depth map based on the first image and the second image of the image set, resulting in a sequence of depth maps. In an example, the image set may include a pair of stereo images, where the first image may be, for example, the left image 722a or 722b while the second image may be, for example, the right image 724a or 724b. The depth maps may include, for example, the depth maps 812. The process 1206 may include determining by the motion detection module, an optical flow based on at least one of, the first images from the sequence of image sets and the second images of the sequence of image sets. The optical flow may be for example, the optical flow 814. The process 1208 may include determining by the motion detection module, motion of an object in the image set based on the optical flow. The process 1210 may include determining TTC with the object based on the sequence of depth maps and the determined motion of the object. The method 1200 may be able to determine TTC using camera images, even when the road lighting is dim. As such, employing the method 1200 on vehicles, may result in avoidance of traffic accidents at night. Various aspects described with respect to the device 800 may be applicable to the method 1200.

[00122] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the process 1204 may include generating a respective unrotated disparity map for each image set using the image processing method 600, and determining the respective depth map based on the respective unrotated disparity map. Determining the depth maps using unrotated disparity maps generated according to the image processing method 600 allows accurate depth to be determined even when the cameras that output the images in the image sets are uncalibrated. The uncalibration

of the cameras may occur as a result of vibrations or movements of the vehicle on which the cameras are mounted.

[00123] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the method 1200 may further include generating a 3D point cloud based on at least one image set of the sequence of image sets, and detecting the object in the three-dimensional point cloud. The 3D point cloud is a dense in data. The 3D point cloud provides detailed information of the surroundings of the vehicle, such that object detection in the 3D point cloud is accurate.

[00124] According to an embodiment which may be combined with any above-described embodiment or with any below described further embodiment, the method 1200 may further include generating a respective three-dimensional point cloud based on each image set of the sequence of image sets, resulting in a sequence of three-dimensional point clouds, comparing distances moved by each point in the sequence of three-dimensional point clouds, and correcting at least one three-dimensional point cloud of the sequence of three-dimensional point clouds, based on the determined distances. By comparing the distances moved by each point, the method 1200 may correct for distortions caused by, for example, camera intrinsics or extrinsics.

[00125] While embodiments of the invention have been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention as defined by the appended claims. The scope of the invention is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced. It will be appreciated that common numerals, used in the relevant drawings, refer to components that serve a similar or the same purpose.

[00126] It will be appreciated to a person skilled in the art that the terminology used herein is for the purpose of describing various embodiments only and is not intended to be limiting of the present invention. 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" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[00127] It is understood that the specific order or hierarchy of blocks in the processes / flowcharts disclosed is an illustration of exemplary approaches. Based upon design preferences, it is understood that the specific order or hierarchy of blocks in the processes / flowcharts may be rearranged. Further, some blocks may be combined or omitted. The accompanying method claims present elements of the various blocks in a sample order, and are not meant to be limited to the specific order or hierarchy presented.

[00128] The previous description is provided to enable any person skilled in the art to practice the various aspects described herein. Various modifications to these aspects will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other aspects. Thus, the claims are not intended to be limited to the aspects shown herein, but is to be accorded the full scope consistent with the language claims, wherein reference to an element in the singular is not intended to mean “one and only one” unless specifically so stated, but rather “one or more.” The word “exemplary” is used herein to mean “serving as an example, instance, or illustration.” Any aspect described herein as “exemplary” is not necessarily to be construed as preferred or advantageous over other aspects. Unless specifically stated otherwise, the term “some” refers to one or more. Combinations such as “at least one of A, B, or C,” “one or more of A, B, or C,” “at least one of A, B, and C,” “one or more of A, B, and C,” and “A, B, C, or any combination thereof” include any combination of A, B, and/or C, and may include multiples of A, multiples of B, or multiples of C. Specifically, combinations such as “at least one of A, B, or C,” “one or more of A, B, or C,” “at least one of A, B, and C,” “one or more of A, B, and C,” and “A, B, C, or any combination thereof” may be A only, B only, C only, A and B, A and C, B and C, or A and B and C, where any such combinations may contain one or more member or members of A, B, or C. All structural and functional equivalents to the elements of the various aspects described throughout this disclosure that are known or later come to be known to those of ordinary skill in the art are expressly incorporated herein by reference and are intended to be encompassed by the claims.

CLAIMS

1. A computer-implemented image processing method (600) comprising:
inputting an image set (212) to a trained machine learning model (104), the image set (212) comprising a first image and a second image;
generating a rotated disparity map (214) of the image set (212), using the trained machine learning model (104);
for each pixel in the first image,
determining, based on the rotated disparity map (214), a rotation angle (426) between the pixel in the first image and its corresponding position in the second image, relative to a common image centre (402), and
determining a respective unrotated disparity based on the respective rotation angle (426); and
generating an unrotated disparity map (216) based on the respective unrotated disparity of each pixel in the first image.
2. The image processing method (600) of any preceding claim, wherein the image set (212) comprises stereo images, wherein the first image is a left stereo image and wherein the second image is a right stereo image.
3. The image processing method (600) of any preceding claim, wherein the trained machine learning model (104) is configured to determine both horizontal disparities and vertical disparities of the image set, such that the rotated disparity map (214) comprises both horizontal disparities and vertical disparities.
4. The image processing method (600) of any preceding claim, wherein a center of the first image is used as the common image center (402).
5. The image processing method (600) of any preceding claim, further comprising:
segmenting the first image into a plurality of first regions;
segmenting the second image into a plurality of second regions corresponding to the plurality of first regions;

for each first region of the plurality of first regions, determining an epipolar line and its direction based on a pixel in the first image and its corresponding position in the second image.

6. The image processing method (600) of claim 5, further comprising:
for each pixel of the first image, determining a respective horizontal disparity based on projection of the respective unrotated disparity through the direction of the epipolar line.
7. The image processing method (600) of any preceding claim, further comprising:
generating a three-dimensional point cloud (208) based on the image set (212) and further based on the unrotated disparity map (216).
8. The image processing method (600) of claim 7, further comprising:
inputting a further image set to the trained machine learning model (104), the further image set comprising a further first image and a further second image captured at a consecutive time frame from the first image and the second image of the image set (212);
generating a further rotated disparity map of the further image set, using the trained machine learning model (104),
for each pixel in the further first image,
identifying a corresponding pixel in the further second image based on the further rotated disparity map, and
determining a respective rotation angle between the pixel in the further first image and the corresponding pixel in the further second image, relative to a further common image centre, and
determining a respective unrotated disparity based on the respective rotation angle;
generating a further unrotated disparity map based on the respective unrotated disparities of each pixel in the further first image;
generating a further three-dimensional point cloud based on the further image set and further based on the further unrotated disparity map;
determine a distance that each point in the three-dimensional point cloud moves from the three-dimensional point cloud (208) to the further three-dimensional point cloud; and

correcting at least one of the three-dimensional point cloud (208) and the further three-dimensional point cloud, based on the determined distances.

9. The image processing method (600) of any preceding claim, wherein the trained machine learning model (104) is trained using a training dataset comprising a plurality of image pairs generated by rotating a pair of calibrated images to a corresponding plurality of different angles, and using a ground truth disparity map (114) generated based on the pair of calibrated images as a training signal.

10. The image processing method (600) of any preceding claim, wherein the trained machine learning model (104) is trained to determine near range disparities, and further separately trained to determine far range disparities.

11. The image processing method (600) of any preceding claim, wherein the trained machine learning model (104) comprises

a feature extraction network (1310) configured to extract features from the image set to generate a feature map (1318), and

wherein the trained machine learning model (104) further comprises a disparity computation network (1320) configured to determine the rotated disparity map (214) based on the generated feature map (1318).

12. The image processing method (600) of claim 11, wherein the feature extraction network (1310) comprises a plurality of neural network branches (1350, 1360), wherein each neural network branch comprises a respective convolutional stack (1312), and a pooling module (1314) connected to the convolutional stack (1312), and wherein the plurality of neural network branches (1350, 1360) share the same weights.

13. The image processing method (600) of any one of claims 11 to 12, wherein the disparity computation network (1320) comprises a three-dimensional convolutional neural network (1324) configured to generate three disparity maps based on the generated feature map (1318).

14. An image processing device (700) comprising:

a processor (702) configured to perform the image processing method (600) of any preceding claim.

15. The image processing device (700) of claim 14, further comprising:
 - a camera (704);
 - an engine (706) for driving the image processing device (700); and/or a steering module (708) for steering the image processing device (700),
 - wherein the processor (702) is configured to use the unrotated disparity maps for driving and/or steering the image processing device (700).
16. A computer-implemented method (1200) for predicting collisions, the method (1200) comprising:
 - inputting a sequence of image sets (720a, 720b) to a motion detection module (802), wherein each image set of the sequence comprises a first image (722a, 722b) and a second image (724a, 724b);
 - determining by the motion detection module (802), for each image set, a respective depth map (812) based on the first image (722a, 722b) and the second image (724a, 724b) of the image set, resulting in a sequence of depth maps (812);
 - determining by the motion detection module (802), an optical flow (814) based on at least one of, the first images (722a, 722b) from the sequence of image sets (720a, 720b) and the second images (724a, 724b) of the sequence of image sets; and
 - determining by the motion detection module (802), motion of an object in the image set based on the optical flow (814); and
 - determining time-to-collision with the object based on the sequence of depth maps (812) and the determined motion of the object.
17. The method (1200) of claim 16, wherein determining by the motion detection module (802), for each image set, a respective depth map (812), comprises generating a respective unrotated disparity map (216) for each image set using the image processing method (600) of claim 1, and determining the respective depth map (812) based on the respective unrotated disparity map (216).
18. The method (1200) of any one of claims 16 to 17, further comprising:

generating a three-dimensional point cloud (208) based on at least one image set of the sequence of image sets (720a, 720b); and
detecting the object in the three-dimensional point cloud (208).

19. The method (1200) of claim 18, further comprising:

generating a respective three-dimensional point cloud (208) based on each image set of the sequence of image sets (720a, 720b), resulting in a sequence of three-dimensional point clouds (208);

comparing distances moved by each point in the sequence of three-dimensional point clouds (208); and

correcting at least one three-dimensional point cloud (208) of the sequence of three-dimensional point clouds (208), based on the determined distances.

20. A device (800) for predicting collisions, the device (800) comprising:

a processor (830) configured to perform the method (1200) of any one of claims 16 to 19.

1 / 8

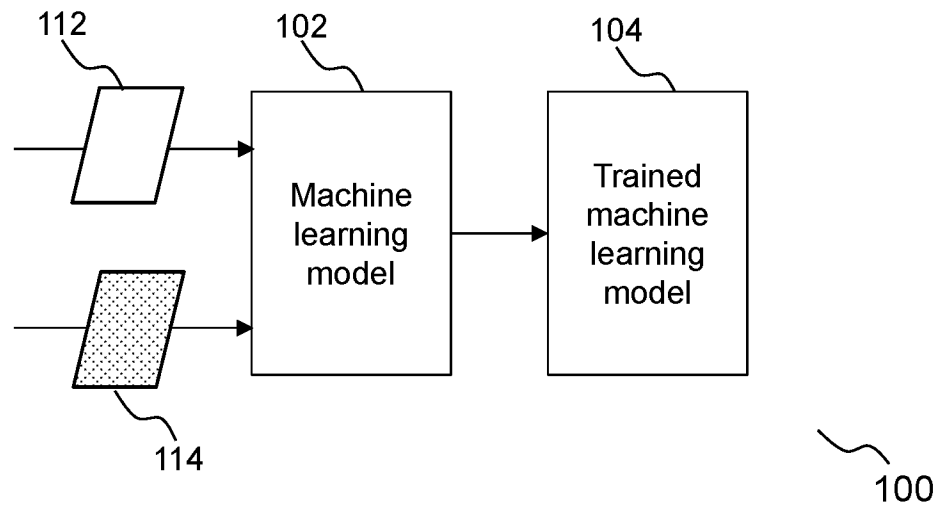


FIG. 1

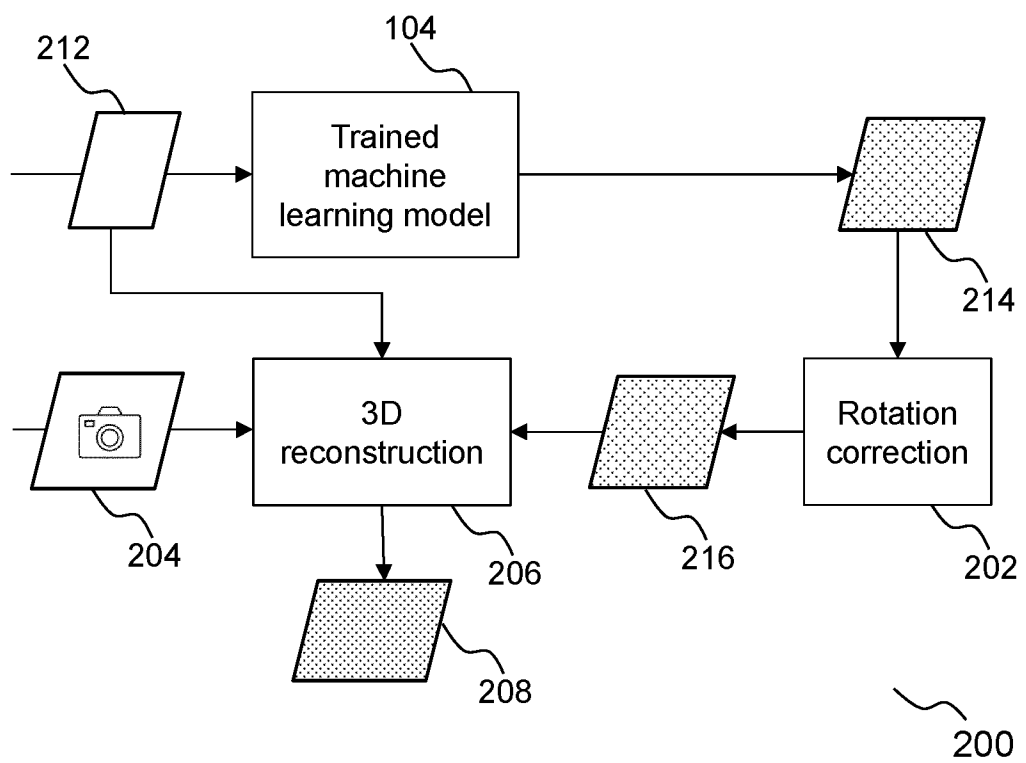


FIG. 2

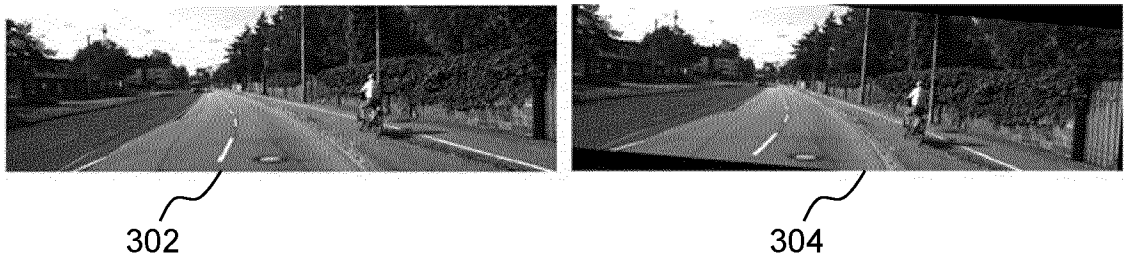


FIG. 3

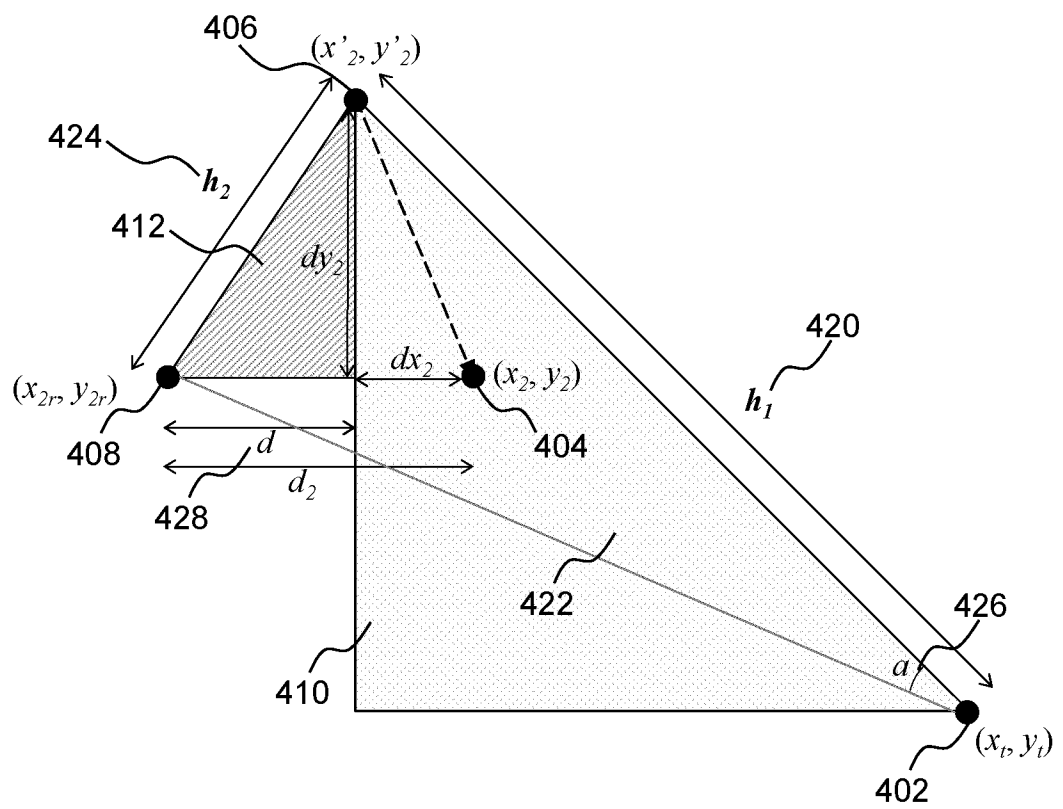


FIG. 4



FIG. 5A

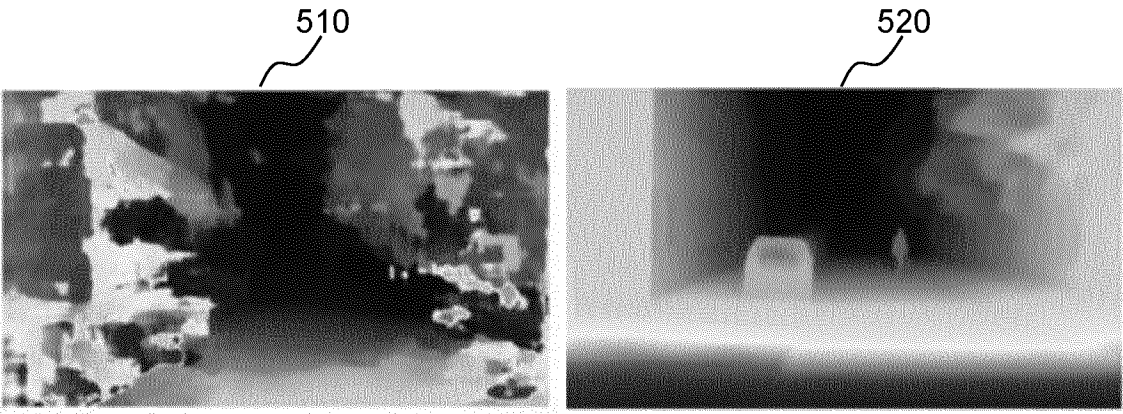


FIG. 5B

4 / 8

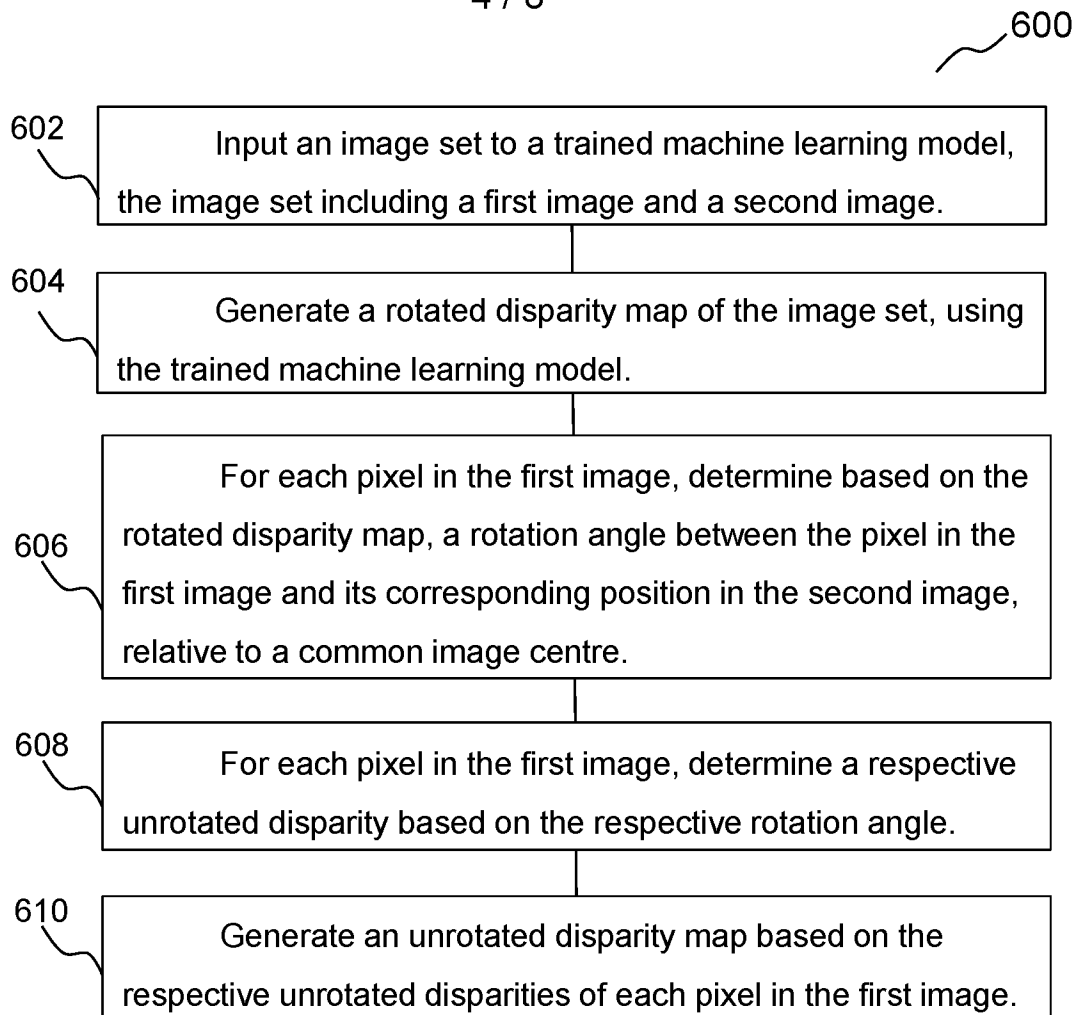


FIG. 6

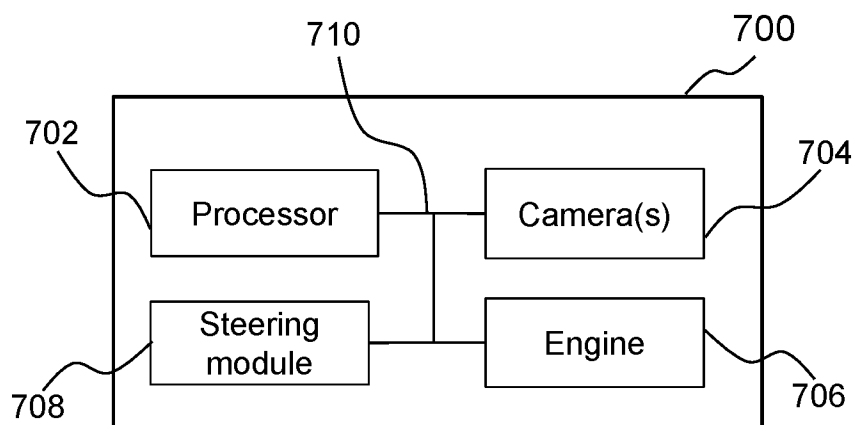


FIG. 7A

5 / 8

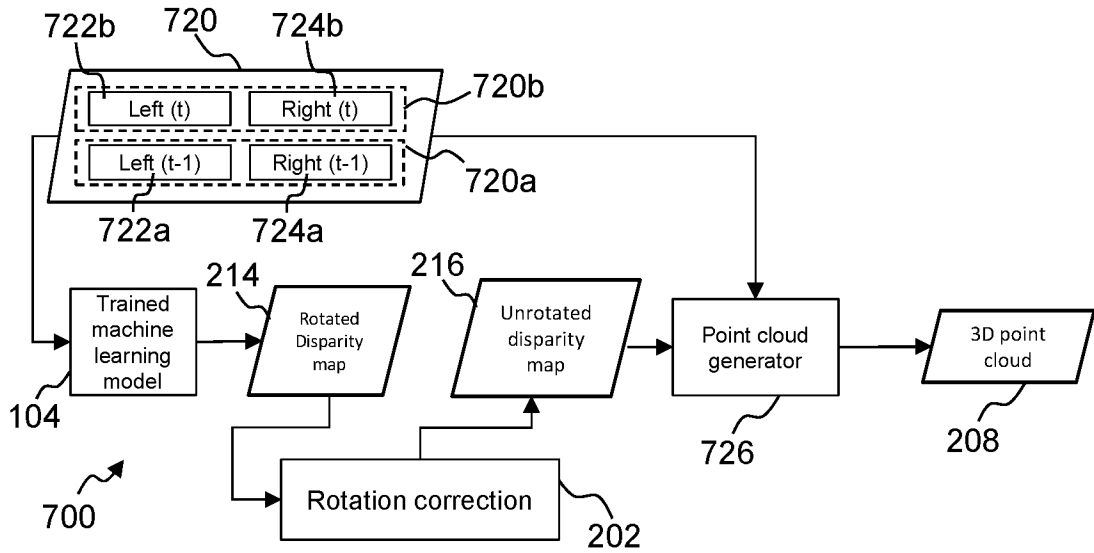


FIG. 7B

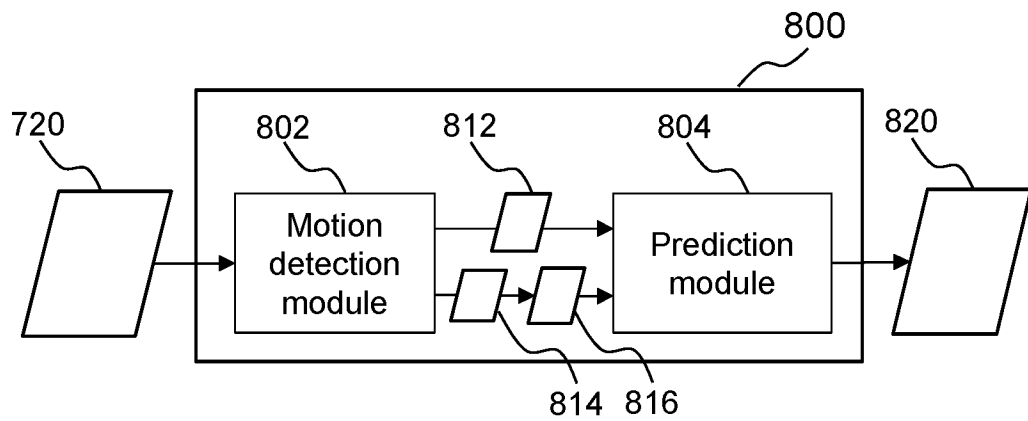


FIG. 8A

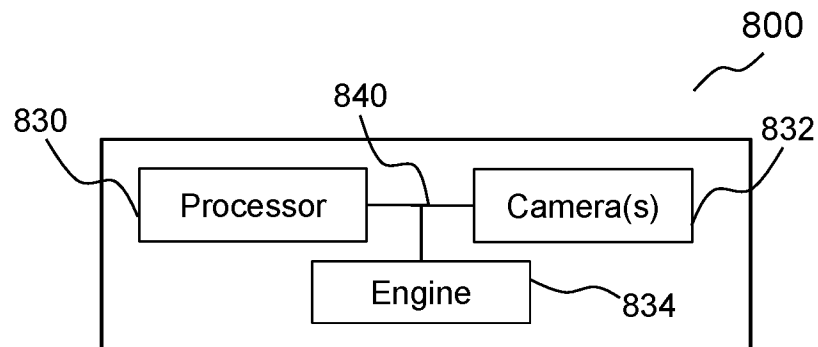


FIG. 8B

6 / 8

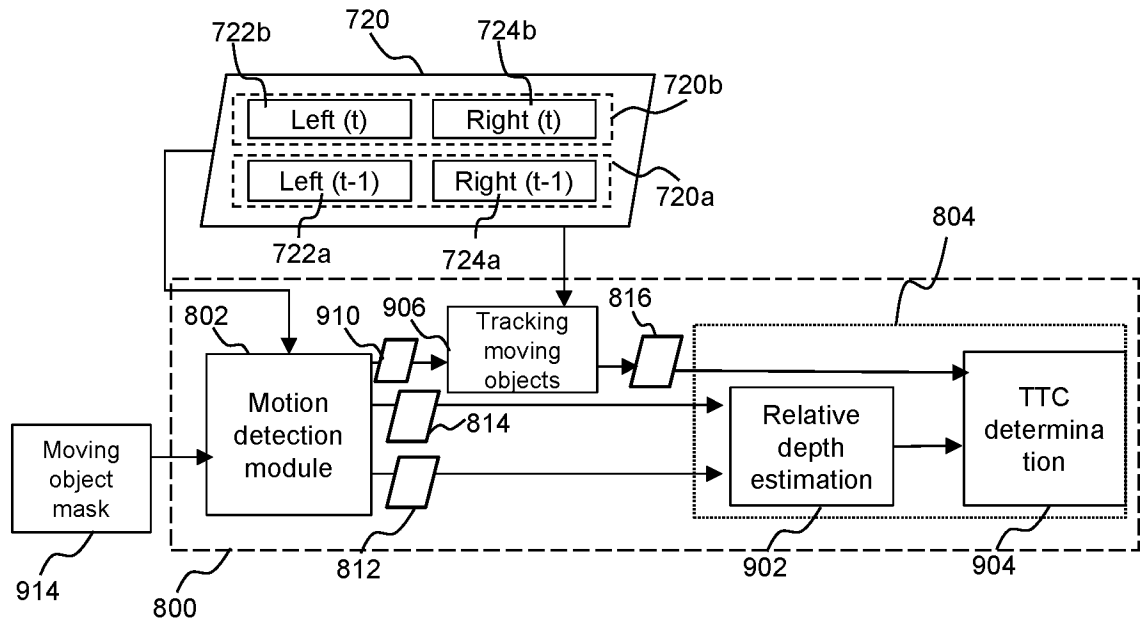


FIG. 9

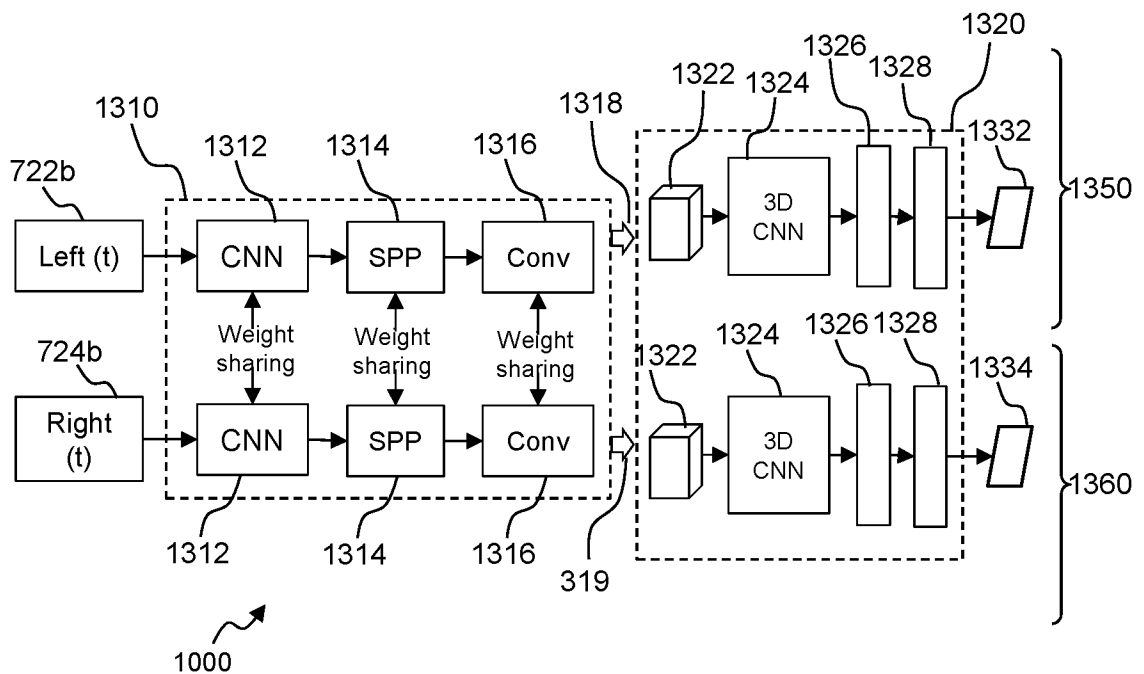


FIG. 10

7 / 8

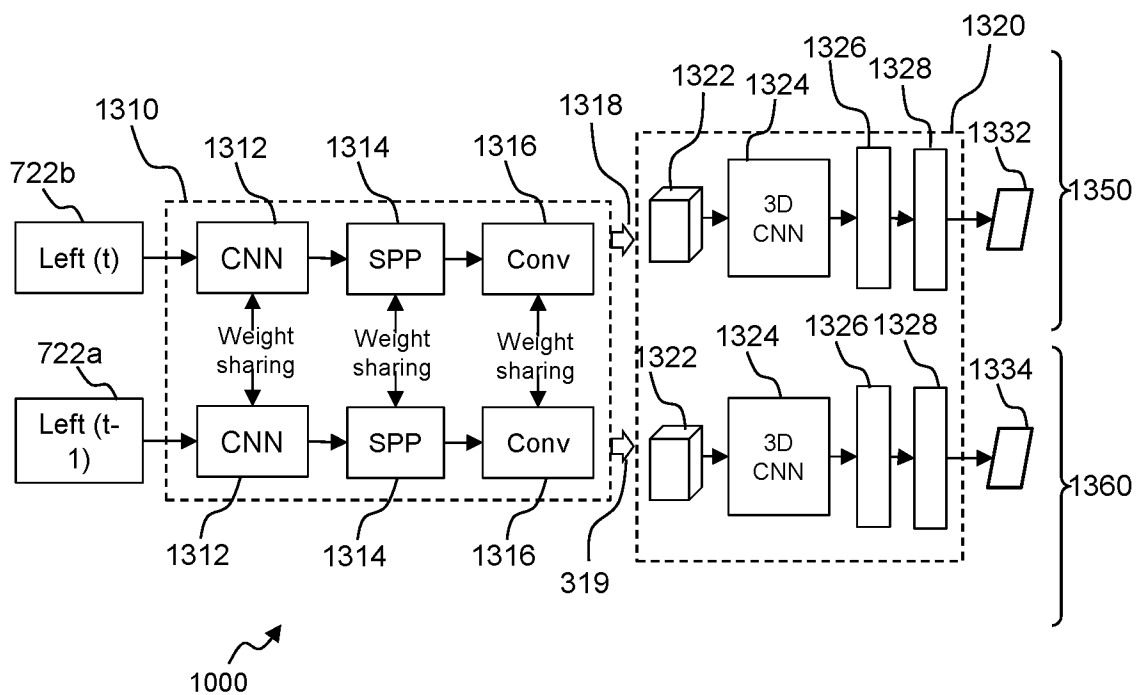


FIG. 11

8 / 8

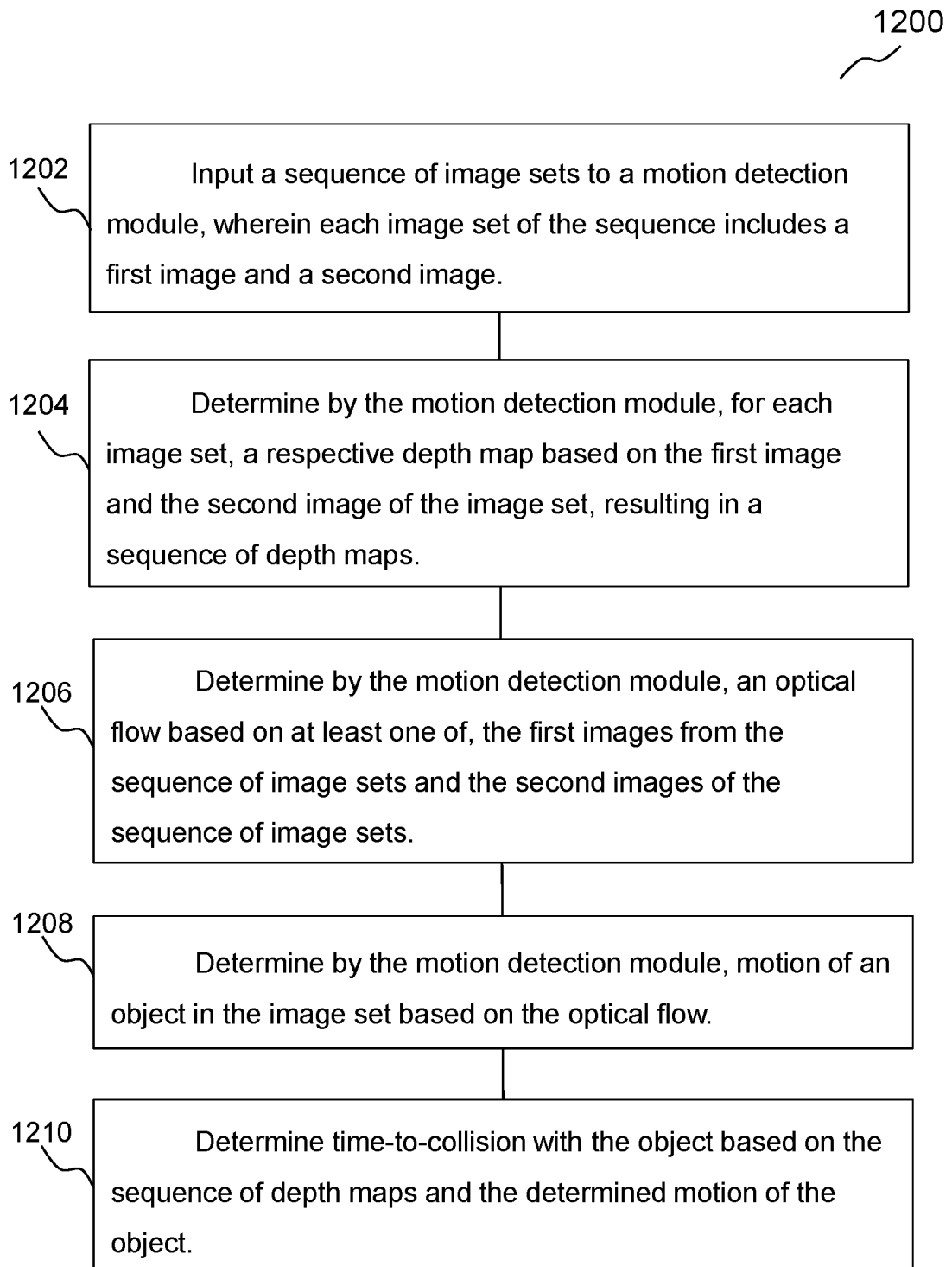


FIG. 12

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2023/079921

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06T7/593 G06V10/82 H04N13/128
ADD.

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)

G06T H04N G06V

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)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113 077 401 B (UNIV ZHEJIANG) 24 June 2022 (2022-06-24)	1-10, 14, 15
Y	the whole document	11-13, 17
Y	CHANG JIA-REN ET AL: "Pyramid Stereo Matching Network", 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, IEEE, 18 June 2018 (2018-06-18), pages 5410-5418, XP033473454, DOI: 10.1109/CVPR.2018.00567 [retrieved on 2018-12-14] Section 3.	11-13



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

"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

9 April 2024

Date of mailing of the international search report

18/04/2024

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Hardell, Alexander

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2023/079921

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ZHANG XUCHONG ET AL: "End-to-end learning of self-rectification and self-supervised disparity prediction for stereo vision", ARXIV,, vol. 494, 26 April 2022 (2022-04-26), pages 308-319, XP087058808, DOI: 10.1016/J.NEUCOM.2022.04.095 [retrieved on 2022-04-26] Section 3. -----	1-15
A	US 2022/309776 A1 (FECHNER THOMAS [DE] ET AL) 29 September 2022 (2022-09-29) paragraphs [0061] - [0065] -----	1-15
X	RATEKE THIAGO ET AL: "Road obstacles positional and dynamic features extraction combining object detection, stereo disparity maps and optical flow data", MACHINE VISION AND APPLICATIONS, SPRINGER VERLAG, DE, vol. 31, no. 7-8, 25 September 2020 (2020-09-25), XP037255193, ISSN: 0932-8092, DOI: 10.1007/S00138-020-01126-W [retrieved on 2020-09-25]	16,18-20
Y	the whole document -----	17-19
Y	YANG WANG ET AL: "Joint Unsupervised Learning of Optical Flow and Depth by Watching Stereo Videos", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 8 October 2018 (2018-10-08), XP081058894, Method - Rigid Alignment Module; figure 3 -----	18,19

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2023/079921

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims;; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☒ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-15

Image processing method and corresponding device, the method including steps of generating a rotated disparity map for an image set comprising a first and a second image using a trained machine learning model, determining a rotation angle and unrotated disparity for each pixel in the first image and its corresponding position in the second image, and generating an unrotated disparity map based on the unrotated disparities for each pixel.

2. claims: 16-20

Method and corresponding device for predicting collisions, the method comprising steps of inputting a sequence of image sets each comprising a first image and a second image into a motion detection module, determining a corresponding sequence of depth maps by the motion detection module, determining an optical flow from the sequence of images, determining motion of an object in the images based on the optical flow, and determining time-to-collision with the object based on the depth maps and the motion of the object.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2023/079921

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CN 113077401	B	24-06-2022	NONE

US 2022309776	A1	29-09-2022	CN 115214681 A 21-10-2022
			DE 102021107904 A1 29-09-2022
			EP 4068223 A1 05-10-2022
			US 2022309776 A1 29-09-2022
