

DETECTION METHOD OF FUSED RADAR INFORMATION WITH CAMERA INFORMATION, AND AUTONOMOUS MOBILE ROBOT FOR THE DETECTION METHOD

BACKGROUND OF THE DISCLOSURE

Technical Field

[0001] The present disclosure relates to an autonomous mobile robot, especially relates to a detection method of the autonomous mobile robot for detecting designated types of objects.

Description of Related Art

[0002] Recently, the development of autonomous mobile robots (AMRs) goes fast, and the requirement for AMRs in the market increases greatly. The AMRs are used in the environments (such as factories, restaurants, or hotels, etc.) full of heterogeneous objects (e.g., human, luggage, tables, or cabinets, etc.). Therefore, how to quickly and accurately detect surrounding environment and objects on the moving path and perform avoidance or corresponding actions, is critically important for the AMRs.

[0003] There are multiple detection manners currently applied to the AMRs, such as Voxel-Net, Voxel-RCNN, Point painting, SFD, and VirConv, etc., with each has different pros and cons. For example, the computation load of SFD is large. VirConv improves SFD and is currently a better detection approach, but it uses all the virtual points detected by the radar for calculations, resulting in creating a considerable computational burden, affecting the AMRs' detection speed, and causing a large amount of energy consumption for the AMRs.

SUMMARY OF THE INVENTION

[0004] The objection of the present disclosure is to provide an autonomous mobile robot and detection method of fused lidar information with camera information, which combining lidar information, camera information, and deep-learning model to implement an accurate detection for designated types of objects. This not only reduces the computation load while detecting, but also improves the security of the AMR due to the increased detection accuracy.

[0005] In one of the exemplary embodiments, the detection method of the present disclosure is incorporated with an autonomous mobile robot (AMR) having a camera, a lidar, and a processor, and includes following steps:

obtaining a camera image of an environment by the camera and obtaining lidar point-cloud information of the environment by the lidar;

obtaining a depth map by the processor based on the camera image, performing a semantic segmentation process on the camera image to obtain semantic scores of each pixel relevant to a designated type, and painting the semantic scores onto the depth map to generate a virtual key instance (VKI), wherein the VKI records a plurality of dense points indicating a target object of the designated type upon a 3D space;

performing a feature extracting procedure on the VKI and the lidar point-cloud information by the processor to obtain multiple VKI features and multiple lidar point-cloud information features;

by the processor, performing a bird's-eye view (BEV) feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon a 2D plane by a virtual instance focused fusion (VIFF) model to generate first bounding box information that corresponds to the target object in the lidar point-cloud information;

performing a region-of-interest (ROI) feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon the 3D space by the VIFF model to generate second bounding box information, wherein the second bounding box information is generated to describe the first bounding box information; and

combining the first bounding box information with the second bounding box information by the VIFF model to output a detection result of describing the target object in the lidar point-cloud information.

[0006] In one of the exemplary embodiments, the AMR of the present disclosure includes:

a camera, configured to obtain a camera image of an environment;

a lidar, configured to obtain lidar point-cloud information of the environment; and

a processor, electrically connected with the camera and the lidar, and configured to execute a plurality of instructions to perform following actions:

obtaining a depth map based on the camera image, performing a semantic segmentation process on the camera image to obtain semantic scores of each pixel relevant to a designated type, and painting the semantic scores onto the depth map to generate a virtual key instance (VKI), wherein the VKI records a plurality of dense points indicating a target object of the designated type upon a 3D space;

performing a feature extracting procedure on the VKI and the lidar point-cloud information to obtain multiple VKI features and multiple lidar point-cloud information features;

performing a BEV feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon a 2D plane by a virtual instance focused fusion (VIFF) model to generate first bounding box information that corresponds to the target object in the lidar point-cloud information;

performing an ROI feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon the 3D space by the VIFF model to generate second bounding box information, wherein the second bounding box information is generated to describe the first bounding box information; and

combining the first bounding box information with the second bounding box information by the VIFF model to output a detection result of describing the target object in the lidar point-cloud information.

[0007] Comparing with Related Art, the present disclosure fuses lidar information with camera information to address traditional problem of poor detection results caused by the sparsity of lidar information. Besides, though modifying the Voxel-based detection approach, the shortcomings of traditional detection manners such as large computational load and slow

detection speed are improved. Also, the present disclosure introduces semantic message while processing, so the detection accuracy as well as the security of the AMRs are further improved.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. 1 is a block diagram of an autonomous mobile robot according to an embodiment of the present disclosure.

[0009] FIG. 2 is a schematic diagram of a detection structure according to an embodiment of the present disclosure.

[0010] FIG. 3 is a flowchart of a detection method according to an embodiment of the present disclosure.

[0011] FIG. 4 is a schematic diagram of a virtual instance focused fusion model according to an embodiment of the present disclosure.

[0012] FIG. 5 is a working flowchart of a BEV feature fusion process according to an embodiment of the present disclosure.

[0013] FIG. 6 is a working flowchart of an ROI feature fusion process according to an embodiment of the present disclosure.

[0014] FIG. 7 is a schematic diagram showing a detection result according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0015] The present disclosure discloses an autonomous mobile robot (AMR) that fuses lidar information with camera information to perform detection. By fusing lidar information with camera information, this AMR addresses the problem of insufficient accuracy caused by the sparsity of lidar information, which occurred when performing the detection solely based on lidar information. Also, the detection method of the present disclosure improves Voxel-based detection method, solving the problem of large computational load and slow detection speed caused by referencing too much information during detection, thereby also reducing the energy consumption of the AMR.

[0016] Besides, the present disclosure introduces semantic messages while processing camera information, thereby further increasing the detection accuracy. Furthermore, by improving the detection accuracy, the security of the AMR is improved accordingly.

[0017] Please refer to FIG. 1, which is a block diagram of an autonomous mobile robot according to an embodiment of the present disclosure. As shown in FIG. 1, the autonomous mobile robot (AMR) 1 of the present disclosure mainly includes a processor 11, a camera 12, and a lidar 13. The processor 11 is electrically connected with the camera 12 and the lidar 13, controls the camera 12 to shot within the environment where the AMR 1 exists to obtain camera information (such as a camera image 121 as shown in FIG. 2), and controls the lidar 13 to perform sensing within the environment to obtain lidar information (such as lidar point-cloud information 131 as shown in FIG. 2). In one of the exemplary embodiments, the lidar 13 can be millimeter wave radar.

[0018] In one of the exemplary embodiments, the processor 11 may be a central processing unit (CPU), a micro control unit (MCU), a programmable logic controller (PLC), a system on chip (SoC), or a field programmable gate array (FPGA), etc. In the present disclosure, the processor 11 executes a plurality of instructions to internally establish a semantic key instance selection (SKIS) model 2 and a virtual instance focused fusion (VIFF) model 3. The SKIS model 2 and the VIFF model 3 are models implemented by software, wherein the SKIS model 2 is configured to generate a plurality of dense points (i.e., a virtual key instance, VKI) for representing an object of a designated type in a three-dimensional (3D) space based on camera information, and the VIFF model 3 is configured to fuse the lidar information with the VKI to generate a detection result of a target object for the designated type (detailed described in the following).

[0019] It should be mentioned that the present disclosure uses the VIFF model 3 to perform fusion on a feature set of the camera information and the lidar information, thereby outputting a detection result that combines the advantages of the camera 12 and the lidar 13.

For example, even if the AMR 1 locates in a dark environment and the camera 12 fails, the lidar 13 does not fail because of darkness, so the technical solution of the present disclosure can still output a reliable detection result. For another example, if the camera 12 is determined to be failure based on the content of the VKI (for example, the VKI is a 0-matrix), the VIFF model 3 can automatically increases the weight of lidar features and reduces the weight of camera features while performing fusion. Therefore, the output detection result will not be unavailable just because the failure of the camera 12.

[0020] As shown in FIG. 1, the AMR 1 of the present disclosure further includes a driving mechanism 14 electrically connected with the processor 11. In some embodiments, the driving mechanism 14 may, for example, include fixed wheels and/or steerable wheels, and the AMR 1 is supported on the ground by the driving mechanism 14. Therefore, the processor 11 can generate corresponding control commands based on the detection result and control the driving mechanism 14 in accordance with the control commands, thereby implementing the movement of the ARM 1.

[0021] As shown in FIG. 1, the AMR 1 of the present disclosure further includes a human-machine interface (HMI) 15 electrically connected with the processor 11. In some embodiments, the HMI 15 may, for example, include physical buttons, touch buttons, touch screen, or voice input interface. Through operating the HMI 15, a user can perform setting to the AMR 1 to issue relevant commands.

[0022] The present disclosure also discloses a detection method of fused radar information with camera information. In the present disclosure, the processor 11 of the AMR 1 can execute a plurality of instructions to internally perform the SKIS model 2 and the VIFF model 3 and then implement each specific step of the detection method described below based on the SKIS model 2 and the VIFF model 3.

[0023] Please refer to FIG. 2 and FIG. 3, wherein FIG. 2 is a schematic diagram of a detection structure according to an embodiment of the present disclosure and FIG. 3 is a flowchart of a detection method according to an embodiment of the present disclosure.

[0024] As shown in FIG. 2, the detection method of the present disclosure continuously obtains camera image 121 and lidar point-cloud information 131 respectively through the camera 12 and the lidar 13 while the AMR 1 moves within the environment. The processor 11 of the AMR 1 uses the SKIS model 2 to process the camera image 121 (and the lidar point-cloud information 131) to generate a virtual key instance (VKI) 4, wherein the VKI 4 records a plurality of dense points that indicates a designated type (such as a human type) of object (such as human) in the environment within a 3D space. Next, the processor 11 of the AMR 1 uses the VIFF model 3 to process the VKI 4 and the lidar point-cloud information 131 to generate a detection result 7 that may indicate an object of the designated type in the lidar point-cloud information 131. In one of the exemplary embodiments, the detection result 7 is a bounding box that can envelops a target object in the lidar point-cloud information, but is not limited thereto.

[0025] In the descriptions below, the structure of FIG. 2 and the flowchart of FIG. 3 are referred together for interpretation.

[0026] First, the AMR 1 moves within the environment through the driving mechanism 14, obtains the camera image 121 of the environment through the camera 12 during its movement, and obtains the lidar point-cloud information 131 of the environment through the lidar 13 during its movement (step S31). Next, the processor 11 processes the camera image 121 to obtain corresponding depth map 22 (step S32). In the present disclosure, the depth map 22 records a group of dense pixel points (for example, 500x1200 pixel points) for representing depth.

[0027] In one of the exemplary embodiments, the camera 12 of the AMR 1 is an RGB camera. In this embodiment, the processor 11 performs a depth completion process 21 on the

camera image 121 and the lidar point-cloud information 131 to generate the depth map 22. In particularly, the processor 11 may include a detection model which is a deep-learning model pre-trained by using data of the designated type (such as the human type). In this embodiment, the processor 11 uses this detection model to perform identification on the camera image 121 to retrieve, from the camera image 121, image belonging to a target object (such as human) of the designated type. Next, the processor 11 performs the depth completion process 21 on the retrieved image and the lidar point-cloud information 131 to generate the depth map 22.

[0028] In another embodiment, the camera 12 of the AMR 1 is an RGB-D camera. In this embodiment, the processor 11 directly retrieves the depth map 22 from the camera image 121 generated by the RGB-D camera without performing the depth completion process 21.

[0029] In another embodiment, the processor 11 further includes a detection model which is a deep-learning model pre-trained by data of the designated type (such as human type). Because the lidar point-cloud information 131 already includes depth information, the processor 11 in this embodiment may, in aforementioned step S32, perform the identification on the lidar point-cloud information 131 through the detection model to directly obtain the depth map 22 corresponding to a target object of the designated type from the lidar point-cloud information 131.

[0030] In sum, the processor 11 of the present disclosure may obtain the depth map 22 solely based on the camera image 121, obtain the depth map 22 solely based on the lidar point-cloud information 131, or obtain the depth map 22 based on the camera image 121 and the lidar point-cloud information 131, but not limited thereto.

[0031] After the step S32, the processor 11 performs a semantic segmentation process 23 on the camera image 121 to obtain semantic scores of each pixel in the camera image 121 relative to the designated type (step S33). In particularly, as mentioned above, the processor 11 includes a detection model for identifying the target object in the camera image 121. Based on the relationship between the target object and each pixel point in the camera image 121 (for

example, no target object exists, a part of the target object exists, and a specified target object exists, etc.), the processor 11 gives a semantic score 25 for each pixel point. Finally, the processor 11, through point painting technique, paints the semantic scores 25 onto the depth map 22 to generate the VKI 4 (step S34). In the present disclosure, the VKI 4 records the plurality of dense points representing the target object of the designated type within the 3D space.

[0032] It is worth saying that after the processor 11 paints the semantic scores 25 onto the depth map 22, painted information 26 can be obtained. Through transforming the painted information 26 into lidar coordinate system (not shown in the FIGs) used by the lidar 13, the processor 11 can obtain the VKI 4.

[0033] In one of the exemplary embodiments, when performing the semantic segmentation process 23 on the camera image 121, the processor 11 uses the detection model to identify the camera image 121 to generate a target object mask 24, wherein the target object mask 24 only includes an image of the object belonging to the designated type being identified by the detection model. Therefore, the processor 11 can use the target object mask 24 to filter the depth map 22 to reduce the complexity of the depth map 22 (i.e., to delete a part of information that does not belong to the target object), thereby reducing the number of points to be computed to improve the computational speed. By using the above filtering procedure, the accuracy of the VKI 4 finally output by the SKIS model 2 can be further optimized.

[0034] Since the light in the environment where the AMR 1 exists may continuously change, in some embodiments, the processor 11 detects the light in the environment (e.g., too dark, normal, or overexposed) and switches at any time to different approaches to generate the depth map 22 and/or the target object mask 24 based on the light conditions (e.g., obtained solely from the camera image 121, obtained solely from the lidar point-cloud information 131, or obtained from a combination of the camera image 121 and the lidar point-cloud information 131).

[0035] Next, the processor 11 respectively performs a feature extracting process 5 on the VKI 4 and the lidar point-cloud information 131 to obtain multiple VKI features and multiple lidar point-cloud information features (step S35). Next, the processor 11 uses the VIFF model 3 to process the multiple VKI features and the multiple lidar point-cloud information features to generate the detection result 7.

[0036] To be more specific, the feature extracting process 5 is a 3D feature extracting procedure that adopts 3D backbone. After respectively performing the 3D feature extracting procedure on the VKI 4 and the lidar point-cloud information 131, the processor 11 can obtain multiple lidar point-cloud information 3D features and multiple VKI 3D features. Also, through performing a dimensionality reduction process 6 on the multiple lidar point-cloud information 3D features and the multiple VKI 3D features, the processor 11 can generate multiple lidar point-cloud information 2D features and multiple VKI 2D features. More specifically, the dimensionality reduction process 6 compresses height information of the multiple lidar point-cloud information 3D features and the multiple VKI 3D features, thereby transforming the 3D features represented in the 3D space (X, Y, Z) into 2D features represented on a 2D plane (X, Y), where the 2D features can also be called the bird's-eye View (BEV) features.

[0037] As shown in FIG. 2, the VIFF model 3 of the present disclosure respectively performs a BEV feature fusion process 31 and a (3D) region-of-interest (ROI) feature fusion process 32 based on software design. The BEV feature fusion process 31 is performed to process the multiple lidar point-cloud information 2D features and the multiple VKI 2D features, while the ROI feature fusion process 32 is performed to process the multiple lidar point-cloud information 3D features and the multiple VKI 3D features.

[0038] To be more specifically, the processor 11 uses the VIFF model 3 to perform the BEV feature fusion process 31 on the multiple VKI features and the multiple lidar point-cloud information features upon a 2D plane (i.e., the multiple VKI 2D features and the multiple lidar point-cloud information 2D features) to generate first bounding box information that is

corresponding to a target object in the lidar point-cloud information 131 (step S36). In one of the exemplary embodiments, the first bounding box information records a rough and unrefined detection box (e.g., without orientation and with fixed length and fixed width), and this detection box roughly envelops the target object in the lidar point-cloud information 131.

[0039] Next, the processor 11 uses the VIFF model 3 to perform the ROI feature fusion process 32 on the multiple VKI features and the multiple lidar point-cloud information features upon a 3D space (i.e., the multiple VKI 3D features and the multiple lidar point-cloud information 3D features) to generate second bounding box information that is used to describe the first bounding box information (step S37). In one of the exemplary embodiments, the second bounding box information records specified information including orientation, length, and width of the detection box.

[0040] Finally, the processor 11 uses the VIFF model 3 to combine the first bounding box information with the second bounding box information to output the detection result 7 that is used to describe the target object in the lidar point-cloud information 131 (step S38).

[0041] As shown in FIG. 2, the VIFF model 3 of the present disclosure performs a two-stages detection. First, the BEV feature fusion process 31 of the VIFF model 3 performs a first stage detection to generate a rough detection box for the target object; next, the ROI feature fusion process 32 of the VIFF model 3 performs a second stage detection to optimize this detection box in orientation, length, and width, etc., and output a detection box (i.e., the detection result 7) that precisely envelops the target object (such as human) in the 3D space.

[0042] Please refer to FIG. 7 at the same time, wherein FIG. 7 is a schematic diagram showing a detection result according to an embodiment of the present disclosure. As shown in FIG. 7, when the AMR 1 moves within the environment, the camera 12 and the lidar 13 respectively obtain the camera image 121 and the lidar point-cloud information 131, and then the SKIS model 2 can identify a target object 8 in the environment (taken human as an example in FIG. 7). Through further processing on the features by the VIFF model 3, the detection result

7 (i.e., the bounding box) that precisely envelops the one or more target objects 8 in the environment can be generated on the lidar point-cloud information 131.

[0043] Please refer to FIG. 4, FIG. 5, and FIG. 6 at the same time, wherein FIG. 4 is a schematic diagram of a virtual instance focused fusion model according to an embodiment of the present disclosure, FIG. 5 is a working flowchart of a BEV feature fusion process according to an embodiment of the present disclosure, and FIG. 6 is a working flowchart of an ROI feature fusion process according to an embodiment of the present disclosure.

[0044] FIG. 4 discloses a structure of the VIFF model 3 of the present disclosure. It can be understood from FIG. 4 that the VIFF model 3 is used to perform two-stages fusion including the BEV feature fusion process 31 and the ROI feature fusion process 32. The BEV feature fusion process 31 is a fusion process performed upon a 2D plane, while the ROI feature fusion process 32 is a process performed upon a 3D space.

[0045] The following description combines FIG. 4 and FIG. 5 to detailed interpret the BEV feature fusion process 31.

[0046] First, as discussed above, after obtaining multiple lidar point-cloud 3D features 511 and multiple VKI 3D features 521 through the feature extracting process 5, the processor 11 needs to first performs the dimensionality reduction process 6 on the multiple lidar point-cloud information 3D features 511 and the multiple VKI 3D features 521 to obtain multiple lidar point-cloud information 2D features 512 and multiple VKI 2D features 522 upon the 2D plane (step S51). Next, the processor 11 regards the multiple lidar point-cloud information 2D features 512 and the multiple VKI 2D features 522 as an input of the BEV feature fusion process 31.

[0047] The BEV feature fusion process 31 concatenates the multiple lidar point-cloud information 2D features 512 with the multiple VKI 2D features 522 along the channel dimension to generate multiple concatenated features, and then performs a channel-attention process 311 on the multiple concatenated features to generate multiple processed features (step

S52). Next, the BEV feature fusion process 31 performs a spatial-attention process 312 on the multiple processed features that passed through the channel-attention process 311 to generate multiple fused BEV features 310 (step S53). Then, the BEV feature fusion process 31 generates the aforementioned first bounding box information based on the multiple fused BEV features 310 (step S54). As discussed above, the first bounding box information records a detection box that roughly envelops the target object in the lidar point-cloud information 131, so the BEV feature fusion process 31 imports the first bounding box information and/or the fused BEV features 310 into the ROI feature fusion process 32 as a 3D region proposal that enables the ROI feature fusion process 32 to identify regions of interest.

[0048] In particularly, the channel-attention process 311 is performed for the model to adjust and learn the importance of different feature channels, thereby knowing which channels within the features are important. The spatial-attention process 312 is performed for complement with the channel-attention process 311 to improve the performance of the model. More specifically, the spatial-attention process 312 uses a heatmap of the VKI 4 to enhance the lidar point-cloud information 2D features 512 and allows the model to adjust and learn the importance of different spatial locations within the features, thereby knowing which spatial locations within the features are important.

[0049] The channel-attention process 311 and the spatial-attention process 312 are common technical solutions in artificial intelligent field, the detailed description will be omitted here.

[0050] The following description combines FIG. 6 and FIG. 5 to further interpret the ROI feature fusion process 32.

[0051] As shown in the FIGs, after obtaining the multiple lidar point-cloud information 3D features 511 and the multiple VKI 3D features 521, the processor 11 directly regards the multiple lidar point-cloud information 3D features 511 and the multiple VKI 3D features 521 as the input of the ROI feature fusion process 32. The ROI feature fusion process 32 respectively

processes the multiple lidar point-cloud information 3D features 511 and the multiple VKI 3D features 521, wherein the process includes performing a first self-attention process 321 on the multiple lidar point-cloud information 3D features 511 to obtain a first vector representation (step S61) and performing a second self-attention process 322 on the multiple VKI 3D features 521 to obtain a second vector representation (step S63).

[0052] In particularly, the ROI feature fusion process 32 of the present disclosure adopts bi-directional self-attention features, which may reduce the noise generated from the frontend depth completion process 21 and compensate for the situation of incomplete calibration between the camera 12 and the lidar 13.

[0053] Next, the ROI feature fusion process 32 performs a first frequency self-attention process 323 on the first vector representation generated from the first self-attention process 321 to generate a first query vector (Q), a first key vector (K), and a first value vector (V) (step S62), and performs a second frequency self-attention process 324 on the second vector representation generated from the second self-attention process 322 to generate a second query vector (Q), a second key vector (K), and a second value vector (V) (step S64).

[0054] In one of the exemplary embodiments, the first frequency self-attention process 323 is a frequency self-attention process focused on low-frequency (Fsal), while the second frequency self-attention process 324 is a frequency self-attention focused on all value (Fsav), but is not limited thereto. The first frequency self-attention process 323 and the second frequency self-attention process 324 separate the frequency domain into different parts (such as high-frequency and low-frequency) and apply different self-attention processes respectively on different parts of features.

[0055] Next, the ROI feature fusion process 32 performs a first cross-attention process 325 on the first value vector and the first key vector generated from the first frequency self-attention process 323 and the second query vector generated from the second frequency self-attention process 324 to generate a third vector representation, and performs a second cross-

attention process 326 on the second value vector and the second key vector generated from the second frequency self-attention process 324 and the first query vector generated from the first frequency self-attention process 323 to generate a fourth vector representation (step S65).

[0056] The cross-attention process allows the model to, while processing features of one sequence, focus on another different sequence at the same time.

[0057] Next, the ROI feature fusion process 32 performs a max pooling process on the third vector representation generated by the first cross-attention process 325, the fourth vector representation generated by the second cross-attention process 326, and the output of the second frequency self-attention process 324 to generate a first output (step S66). To be more specific, the max pooling process is performed to retain the most prominent, active information at each position in the feature representations from different processing paths.

[0058] Finally, the ROI feature fusion process 32 concatenates the first output of the max pooling process with the output of the first frequency self-attention process 323 to generate fused three-dimensional (3D) ROI features 320 (step S67). Further, the ROI feature fusion process 32 generates a second bounding box information based on the fused BEV features 310 of the BEV feature fusion process 31 and/or the first bounding box information and the fused 3D ROI features 320 (step S68). Finally, as shown in the step S38 of FIG. 3, the processor 11 generates the detection result 7 indicating the target object in the lidar point-cloud information 131 based on the first bounding box information and the second bounding box information.

[0059] The aforementioned self-attention process, frequency self-attention process, cross-attention process, max pooling process, query vector, key vector, and value vector are common technique used in AI field, the detailed descriptions are omitted here.

[0060] As discussed above, the AMR 1 of the present disclosure obtains the camera image 121 through the camera 12 and obtains the lidar point-cloud information 131 through the lidar 13, and the processor 11 executes a plurality of instructions to implement each specified

step of the detection method of the present disclosure. In particular, the processor 11 executes the plurality of instructions to perform actions listed below in order:

[0061] Action 1: obtaining the depth map 22 based on the camera image 121, performing the semantic segmentation process 23 on the camera image 121 to obtain semantic scores 25 of each pixel relevant to the designated type, and painting the semantic scores 25 onto the depth map 22 to generate the VKI 4, wherein the VKI 4 records a plurality of dense points that indicate the target object of the designated type upon the 3D space.

[0062] Action 2: performing the feature extracting procedure on the VKI 4 and the lidar point-cloud information 131 to obtain multiple VKI features and multiple lidar point-cloud information features.

[0063] Action 3: performing the BEV feature fusion process 31 on the multiple VKI features and the multiple lidar point-cloud information features by the VIFF model 3 upon the 2D plane to generate the first bounding box information corresponding to the target object in the lidar point-cloud information 131.

[0064] Action 4: performing the ROI feature fusion process 32 on the multiple VKI features and the multiple lidar point-cloud information features by the VIFF model 3 upon the 3D space to generate the second bounding box information, wherein the second bounding box information is used to describe the first bounding box information.

[0065] Action 5: combining the first bounding box information with the second bounding box information by the VIFF model 3 to output the detection result 7 that describes the target object in the lidar point-cloud information 131.

[0066] In particularly, through executing the plurality of instructions, the AMR 1 of the present disclosure may implement each specified step as shown in the flowcharts of FIG. 3, FIG. 5, and FIG. 7.

[0067] The applicant, based on JRDB dataset, performed average precision (AP) tests on the detection method of the present disclosure, Voxel R-CNN, and VirConv-L, using

intersection over union (IoU) thresholds of 0.5, 0.25, and 0.25 for the simple, medium, and hard modes (the IoU thresholds for the hard mode is the same as the medium mode, with the difference in difficulty coming from the characteristics of the objects themselves). After the experimental tests, the testing results are shown below in table one and table two.

Algorithm	3D AP			Computational Time(ms)
	Simple	Medium	Hard	
Voxel R-CNN	41.6	50.9	50.5	46
VirConv-L	43.5	58.1	51.5	116
Present Disclosure	43.4	57.9	58.2	52

Table one

[0068] The above table one shows results of 3D pedestrian detection on JRDB validation set. From table one above, it can be seen that the accuracy and the computational time of the algorithm in the present disclosure are significantly better than VirConv-L. Besides, although the computational time of the algorithm in the present disclosure is still slower than Voxel R-CNN, its accuracy in simple, medium, and hard modes is superior to Voxel R-CNN.

Algorithm	BEV AP			Computational Time(ms)
	Simple	Medium	Hard	
Voxel R-CNN	58.2	69.6	61.9	46
VirConv-L	58.9	70.2	62.3	116
Present Disclosure	59.0	70.1	70.4	52

Table two

[0069] The above table two shows results of BEV pedestrian detection on JRDB validation set. From the table two, it can be seen that the accuracy and computational time of the application in the present disclosure are superior to VirConv-L. Besides, although the

computational time of the algorithm in the present disclosure is still slower than Voxel R-CNN, its accuracy in simple, medium, and hard modes is superior to Voxel R-CNN.

[0070] The technical solution of the present disclosure addresses the problem of poor detection results for target objects (such as human) by traditional AMRs by fusing lidar information with camera information. Furthermore, by using the SKIS model and the VIFF model, the present disclosure reduces the amount of information, thereby improving the shortcomings of large computational load and slow detection speed from traditional detection. Moreover, the SKIS model of the present disclosure introduces semantic messages while processing, so as to further improve the detection accuracy.

WHAT IS CLAIMED IS:

1. A detection method of fused lidar information with camera information incorporated with an autonomous mobile robot (AMR) having a camera, a lidar, and a processor, comprising:

step a) obtaining a camera image of an environment by the camera and obtaining lidar point-cloud information of the environment by the lidar;

step b) obtaining a depth map by the processor based on the camera image, performing a semantic segmentation process on the camera image to obtain semantic scores of each pixel relevant to a designated type, and painting the semantic scores onto the depth map to generate a virtual key instance (VKI), wherein the VKI records a plurality of dense points indicating a target object of the designated type upon a 3D space;

step c) performing a feature extracting procedure on the VKI and the lidar point-cloud information by the processor to obtain multiple VKI features and multiple lidar point-cloud information features;

step d) by the processor, performing a bird's-eye view (BEV) feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon a 2D plane by a virtual instance focused fusion (VIFF) model to generate first bounding box information that corresponds to the target object in the lidar point-cloud information;

step e) performing a region-of-interest (ROI) feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon the 3D space by the VIFF model to generate second bounding box information, wherein the second bounding box information is generated to describe the first bounding box information; and

step f) combining the first bounding box information with the second bounding box information by the VIFF model to output a detection result of describing the target object in the lidar point-cloud information.

2. The detection method of claim 1, wherein the designated type is a human type and the target object is human.

3. The detection method of claim 1, wherein the step b) comprises performing a depth completion process on the camera image and the lidar point-cloud information by the processor to generate the depth map, or directly retrieving, by the processor, the depth map from the camera image generated by an RGB-D camera.

4. The detection method of claim 1, wherein the step b) comprises performing the semantic segmentation process on the camera image by the processor to generate a target object mask and filtering the depth map through the target object mask to reduce complexity of the depth map, wherein the processor adopts a deep-learning model pre-trained by data of the designated type to perform the semantic segmentation process.

5. The detection method of claim 4, wherein the step b) further comprises painting the semantic scores onto the depth map by the processor to generate painted information and transforming the painted information to a lidar coordinate system used by the lidar to generate the VKI.

6. The detection method of claim 1, wherein the step c) comprises performing a 3D feature extracting procedure on the lidar point-cloud information and the VKI by the processor to obtain multiple lidar point-cloud information 3D features and multiple VKI 3D features, wherein in the step e), the VIFF model performs the ROI feature fusion process on the multiple lidar point-cloud information 3D features and the multiple VKI 3D features.

7. The detection method of claim 6, further comprising a step c1) after the step c): performing a dimensionality reduction process on the multiple lidar point-cloud information 3D features and the multiple VKI 3D features by the processor to generate multiple lidar point-cloud information 2D features and multiple VKI 2D features, wherein in the step d), the VIFF model performs the BEV feature fusion process on the multiple lidar point-cloud information 2D features and the multiple VKI 2D features.

8. The detection method of claim 7, wherein the step d) comprises following steps to implement the BEV feature fusion process:

step d1) concatenating the multiple lidar point-cloud information 2D features with the multiple VKI 2D features along channel dimension to generate multiple concatenated features, and performing a channel-attention process on the multiple concatenated features to generate multiple processed features;

step d2) performing a spatial-attention process on the multiple processed features to generate multiple fused BEV features; and

step d3) generating the first bounding box information based on the multiple fused BEV features.

9. The detection method of claim 8, wherein the step e) comprises following steps to implement the ROI feature fusion process:

step e1) performing a first self-attention process on the multiple lidar point-cloud information 3D features to obtain a first vector representation;

step e2) performing a first frequency self-attention process on the first vector representation to generate a first query vector (Q), a first key vector (K), and a first value vector (V);

step e3) performing a second self-attention process on the multiple VKI 3D features to obtain a second vector representation;

step e4) performing a second frequency self-attention process on the second vector representation to generate a second query vector (Q), a second key vector (K), and a second value vector (V);

step e5) performing a first cross-attention process on the first value vector, the first key vector, and the second query vector to generate a third vector representation, and performing a second cross-attention process on the second value vector, the second key vector, and the first query vector to generate a fourth vector representation;

step e6) performing a max pooling process on the third vector representation, the fourth vector representation, and an output of the second frequency self-attention process to generate a first output;

step e7) concatenating the first output with an output of the first frequency self-attention process to generate fused 3D ROI features; and

step e8) generating the second bounding box information based on the fused BEV features and the fused 3D ROI features.

10. The detection method of claim 9, wherein the first frequency self-attention process is a frequency self-attention process focused on low-frequency (F_{sal}), and the second frequency self-attention process is a frequency self-attention focused on all value (F_{sav}).

11. An autonomous mobile robot (AMR), comprising:

a camera, configured to obtain a camera image of an environment;

a lidar, configured to obtain lidar point-cloud information of the environment; and

a processor, electrically connected with the camera and the lidar, and configured to execute a plurality of instructions to perform following actions:

action a) obtaining a depth map based on the camera image, performing a semantic segmentation process on the camera image to obtain semantic scores of each pixel relevant to a designated type, and painting the semantic scores onto the depth map to generate a virtual key instance (VKI), wherein the VKI records a plurality of dense points indicating a target object of the designated type upon a 3D space;

action b) performing a feature extracting procedure on the VKI and the lidar point-cloud information to obtain multiple VKI features and multiple lidar point-cloud information features;

action c) performing a BEV feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon a 2D plane by a virtual instance focused fusion (VIFF) model to generate first bounding box information that corresponds to the target object in the lidar point-cloud information;

action d) performing an ROI feature fusion process on the multiple VKI features and the multiple lidar point-cloud information features upon the 3D space by the VIFF model to generate second bounding box information, wherein the second bounding box information is generated to describe the first bounding box information; and

action e) combining the first bounding box information with the second bounding box information by the VIFF model to output a detection result of describing the target object in the lidar point-cloud information.

12. The AMR of claim 11, wherein the designated type is a human type and the target object is human.

13. The AMR of claim 11, wherein the action a) comprises performing a depth completion process on the camera image and the lidar point-cloud information to generate the depth map, or directly retrieving the depth map from the camera image generated by an RGB-D camera.

14. The AMR of claim 11, wherein the action a) comprises performing the semantic segmentation process on the camera image to generate a target object mask and filtering the depth map through the target object mask to reduce complexity of the depth map, wherein the processor is configured to adopt a deep-learning model pre-trained by data of the designated type to perform the semantic segmentation process.

15. The AMR of claim 14, wherein the action a) further comprises painting the semantic scores onto the depth map to generate painted information and transforming the painted information to a lidar coordinate system used by the lidar to generate the VKI.

16. The AMR of claim 11, wherein the action b) comprises performing a 3D feature extracting procedure on the lidar point-cloud information and the VKI to obtain multiple lidar point-cloud information 3D features and multiple VKI 3D features, wherein in the action d), the VIFF model performs the ROI feature fusion process on the multiple lidar point-cloud information 3D features and the multiple VKI 3D features.

17. The AMR of claim 16, wherein the processor is configured to further perform an action b1) after the action b): performing a dimensionality reduction process on the multiple lidar point-cloud information 3D features and the multiple VKI 3D features to generate multiple lidar point-cloud information 2D features and multiple VKI 2D features, wherein in the action c), the VIFF model performs the BEV feature fusion process on the multiple lidar point-cloud information 2D features and the multiple VKI 2D features.

18. The AMR of claim 17, wherein the action c) comprises following actions to implement the BEV feature fusion process:

action c1) concatenating the multiple lidar point-cloud information 2D features with the multiple VKI 2D features along channel dimension to generate multiple concatenated features, and performing a channel-attention process on the multiple concatenated features to generate multiple processed features;

action c2) performing a spatial-attention process on the multiple processed features to generate multiple fused BEV features; and

action c3) generating the first bounding box information based on the multiple fused BEV features.

19. The AMR of claim 18, wherein the action d) comprises following actions to implement the ROI feature fusion process:

action d1) performing a first self-attention process on the multiple lidar point-cloud information 3D features to obtain a first vector representation;

action d2) performing a first frequency self-attention process on the first vector representation to generate a first query vector, a first key vector, and a first value vector;

action d3) performing a second self-attention process on the multiple VKI 3D features to obtain a second vector representation;

action d4) performing a second frequency self-attention process on the second vector representation to generate a second query vector, a second key vector, and a second value vector;

action d5) performing a first cross-attention process on the first value vector, the first key vector, and the second query vector to generate a third vector representation, and performing a second cross-attention process on the second value vector, the second key vector, and the first query vector to generate a fourth vector representation;

action d6) performing a max pooling process on the third vector representation, the fourth vector representation, and an output of the second frequency self-attention process to generate a first output;

action d7) concatenating the first output with an output of the first frequency self-attention process to generate fused 3D ROI features; and

action d8) generating the second bounding box information based on the fused BEV features and the fused 3D ROI features.

20. The AMR of claim 19, wherein the first frequency self-attention process is a frequency self-attention process focused on low-frequency (F_{sal}), and the second frequency self-attention process is a frequency self-attention focused on all value (F_{sav}).