

DEVICE AND METHOD FOR TRACKING HUMAN SUBJECTS

Cross Reference to Related Application(s)

- 5 The present disclosure claims the benefit of Singapore Patent Application No. 10201807263Q filed on 27 August 2018, which is incorporated in its entirety by reference herein.

Technical Field

10

The present disclosure generally relates to tracking of human subjects. More particularly, the present disclosure describes various embodiments of a device and a computerized method for tracking the human subjects, i.e. people or persons.

15 Background

Various applications rely or use data on positions of people over time, specifically for detecting and tracking people. Such applications are commonly employed by many assistive and service robotic devices or robots, which are growing in numbers
20 especially in populated areas. These applications often have close interactions with people and as such, detection and tracking of people are fundamental qualities in people-to-robot interactions such as to recognize human activities, attributes, and social relations. These interactions allow robots to understand the intentions of people who can benefit from actions of the robots. Improving these interactions can
25 lead to future development of assistive and service robots.

Significant progress has been made in developing algorithms for detecting and tracking people, often with the aim of improving people-robot interactions or robot navigation abilities in populated areas. For example, vision-based methods have
30 achieved promising performance in recent years with the development of deep learning. Particularly, R-CNN (region-based convolutional neural network) algorithms and variants thereof have demonstrated improvement in detecting and tracking people. However, current vision-based methods still suffer from problems such as an

overwhelming number of possible objects in the robot's field of view. This problem leads to accumulation of errors in detecting and tracking people especially after tracking over an extended time period. These errors may be further exacerbated by partial or full occlusions and rapid scaling or appearance changes of objects in the robot's field of view. Moreover, when deep learning algorithms are implemented, due to the overwhelming number of possible objects, the computation resource requirements and power consumption are increased, and therefore would not be suitable for robotic applications where power resources are constrained or limited.

Therefore, in order to address or alleviate at least one of the aforementioned problems and/or disadvantages, there is a need to provide an improved device and method for tracking people or human subjects.

Summary

According to a first aspect of the present disclosure, there is a device for tracking human subjects, the device comprising a 3D depth tracker, a 2D laser tracker, and a track fusion module. The 3D depth tracker is configured for: capturing 3D mapping data of a space containing one or more human subjects; constructing 3D representations of the space from the 3D mapping data; and generating a first track for each human subject in each 3D representation, the first tracks for tracking the respective human subject. The 2D laser tracker is configured for: capturing 2D mapping data of the space; constructing 2D representations of the space from the 2D mapping data; and generating a second track for each human subject in each 2D representation, the second tracks for tracking the respective human subject. The track fusion module is configured for: associating the respective first tracks with the respective second tracks for each human subject; and collaboratively tracking each human subject based on the respective associated first and second tracks.

According to a second aspect of the present disclosure, there is a method for tracking human subjects. The method comprises: capturing 3D mapping data of a space using a 3D depth tracker, the space containing one or more human subjects; constructing 3D representations of the space from the 3D mapping data; generating

a first track for each human subject in each 3D representation, the first tracks for tracking the respective human subject; capturing 2D mapping data of the space using a 2D laser tracker; constructing 2D representations of the space from the 2D mapping data; generating a second track for each human subject in each 2D representation, the second tracks for tracking the respective human subject; associating the respective first tracks with the respective second tracks for each human subject; and collaboratively tracking each human subject based on the respective associated first and second tracks.

An advantage of the present disclosure is that the different tracking characteristics of the 3D depth tracker and 2D laser tracker complement each other for collaborative tracking of the human subjects. Such collaborative tracking can achieve better tracking results in terms of reliability and accuracy, compared to tracking the human subjects using either the 3D depth tracker and 2D laser tracker alone.

A device and method for tracking human subjects according to the present disclosure are thus disclosed herein. Various features, aspects, and advantages of the present disclosure will become more apparent from the following detailed description of the embodiments of the present disclosure, by way of non-limiting examples only, along with the accompanying drawings.

Brief Description of the Drawings

Figure 1 is an illustration of a device for tracking human subjects.

Figure 2 is a flowchart illustration of a method for tracking human subjects.

Figure 3 is a flowchart illustration of a process performed by a 3D depth tracker of the device for tracking the human subjects.

Figure 4A is an illustration of a space containing the human subjects captured by the 3D depth tracker.

Figure 4B is a flowchart illustration of a process performed by the 3D depth tracker for constructing stixel representations and detecting proposals of the human subjects.

Figure 4C is a flowchart illustration of another process performed by the 3D depth tracker for constructing stixel representations and detecting proposals of the human subjects.

Figure 4D is an illustration of the proposals which are aligned.

Figure 5 is a flowchart illustration of a process performed by a 2D laser tracker of the device.

Figure 6A and Figure 6B are illustrations of a spatial map of a robotic device for tracking and following a target person.

Figure 7 is a flowchart illustration of a process performed by the robotic device for calculating velocities to follow the target person.

Detailed Description

For purposes of brevity and clarity, descriptions of embodiments of the present disclosure are directed to a device and method for tracking human subjects, in accordance with the drawings. While aspects of the present disclosure will be described in conjunction with the embodiments provided herein, it will be understood that they are not intended to limit the present disclosure to these embodiments. On the contrary, the present disclosure is intended to cover alternatives, modifications and equivalents to the embodiments described herein, which are included within the scope of the present disclosure as defined by the appended claims. Furthermore, in the following detailed description, specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be recognized by an individual having ordinary skill in the art, i.e. a skilled person, that the present disclosure may be practiced without specific details, and/or with multiple details arising from combinations of aspects of particular embodiments. In a number of

instances, well-known systems, methods, procedures, and components have not been described in detail so as to not unnecessarily obscure aspects of the embodiments of the present disclosure.

- 5 In embodiments of the present disclosure, depiction of a given element or consideration or use of a particular element number in a particular figure or a reference thereto in corresponding descriptive material can encompass the same, an equivalent, or an analogous element or element number identified in another figure or descriptive material associated therewith.

10

References to “an embodiment / example”, “another embodiment / example”, “some embodiments / examples”, “some other embodiments / examples”, and so on, indicate that the embodiment(s) / example(s) so described may include a particular feature, structure, characteristic, property, element, or limitation, but that not every
15 embodiment / example necessarily includes that particular feature, structure, characteristic, property, element or limitation. Furthermore, repeated use of the phrase “in an embodiment / example” or “in another embodiment / example” does not necessarily refer to the same embodiment / example.

- 20 The terms “comprising”, “including”, “having”, and the like do not exclude the presence of other features / elements / steps than those listed in an embodiment. Recitation of certain features / elements / steps in mutually different embodiments does not indicate that a combination of these features / elements / steps cannot be used in an embodiment.

25

As used herein, the terms “a” and “an” are defined as one or more than one. The use of “/” in a figure or associated text is understood to mean “and/or” unless otherwise indicated. The term “set” is defined as a non-empty finite organization of elements that mathematically exhibits a cardinality of at least one (e.g. a set as defined herein
30 can correspond to a unit, singlet, or single-element set, or a multiple-element set), in accordance with known mathematical definitions. The recitation of a particular numerical value or value range herein is understood to include or be a recitation of an approximate numerical value or value range. As used herein, the terms “first” and

“second” are used merely as labels or identifiers and are not intended to impose numerical requirements on their associated terms.

5 In representative or exemplary embodiments of the present disclosure, there is a device 100 configured for performing a method 200 for tracking human subjects. With reference to Figure 1, the device 100 includes a three-dimensional (3D) depth tracker 120 and a two-dimensional (2D) laser tracker 140. The device 100 further includes a processor for performing various steps of the method 200. For example, the processor cooperates with various modules / components of the device 100,
10 such as a track fusion module 160 and a pursuit controller module 180.

The method 200 for tracking human subjects is described with reference to Figure 2. Tracking of human subjects includes firstly using a detection module to detect the human subjects to thereby acquire detection data. A recognition module processes
15 the acquired detection data to recognize the human subjects to be tracked. A tracking module is then used to track the recognized human subjects as the intended targets.

In computer vision, the objective of detection is to notice or discover the presence of
20 an object within a field of view of a tracker, specifically within an image or video frame captured by the tracker. Various algorithms and methods can be used to detect objects. For example, edge detection methods such as Canny edge detection can be used to detect objects in an image by defining object boundaries which can be found by looking at intensity variations across the image. Other detection
25 methods are used to detect objects in a video, including background subtraction methods to create foreground masks, such as Gaussian mixture model (GMM) and absolute difference model. Object recognition is a process for identifying or knowing the nature of an object in an image or video frame. Recognition can be based on matching, learning, or pattern recognition algorithms with the objective being to
30 classify an object. Various algorithms and methods, such as matching, learning, or pattern recognition algorithms, can be used to recognize objects with the objective being to classify the objects. Machine learning algorithms can be trained with training data to improve object recognition. The objective of object tracking then is to keep

watch on a target object by the path of the object in successive images or video frames. Particularly, tracking algorithms are used to locate and watch on one or more moving objects over time in a video stream.

- 5 The method 200 includes a step 202 of capturing 3D mapping data of a space using the 3D depth tracker 120. The space refers to an environment which the 3D depth tracker 120 is capturing and the space contains one or more human subjects. The 3D mapping data are captured continually over a time period and at discrete times within the time period, such that each 3D representation is associated with one of the
10 discrete times. The method 200 includes a step 204 of constructing 3D representations of the space from the 3D mapping data. The method 200 includes a step 206 of generating a first track for each human subject in each 3D representation, the first tracks for tracking the respective human subject. Without being limiting, the first track may be taken to mean an imprint or trace of the human subject in the 3D
15 representation at the respective discrete time. A human subject can thus be tracked over the time period using the first tracks generated for each discrete time within the time period.

The method 200 includes a step 208 of capturing 2D mapping data of the space
20 using the 2D laser tracker 140. The method 200 includes a step 210 of constructing 2D representations of the space from the 2D mapping data. The method 200 includes a step 212 of generating a second track for each human subject in each 2D representation, the second tracks for tracking the respective human subject. Without being limiting, the second track may be taken to mean an imprint or trace of the
25 human subject in the 2D representation at the respective discrete time. A human subject can thus be tracked over the time period using the second tracks generated for each discrete time within the time period.

Although the terms “first track” and “second track” have been used to identify the
30 tracks generated by the 3D depth tracker 120 and 2D laser tracker 140, respectively, the terms “first” and “second” are merely labels and do not impose numerical requirements on the tracks. In other words, it will be appreciated that the tracks

determined by the 3D depth tracker 120 and 2D laser tracker 140 may be reversely labelled or identified as “second track” and “first track”, respectively.

5 The method 200 includes a step 214 performed by the track fusion module 160 of the device 100 to associate the respective first tracks with the respective second tracks for each human subject. The method 200 includes a step 216 performed by the track fusion module 160 to collaboratively track each human subject based on the respective associated first and second tracks. Said collaborative tracking of the human subjects is achieved by associating and fusing data from the respective first
10 and second tracks, such as by use of a probabilistic aggregation scheme. Said fusing of data may be referred to as sensor fusion which is defined as the combining of sensory data or data derived from disparate sources, such that the resulting information has less uncertainty than would be possible when these sources were used individually. The reduction in uncertainty means that the collaborative tracking
15 using the associated first and second tracks is more accurate and reliable than using individual first and second tracks.

The 3D depth tracker 120 is configured to perform at least the steps 202, 204, and 206 of the method 200. The 3D depth tracker 120 includes a depth sensor such as
20 those used in Microsoft Kinect, Asus Xtion, and PrimeSense Carmine. The depth sensor typically includes an infrared projector and an infrared camera cooperating together. The infrared projector projects a pattern of infrared illumination on an object. The infrared pattern reflects from the object and is captured by the infrared camera. The depth sensor processes the captured infrared pattern and computes depth data
25 from the displacement of the infrared pattern. The infrared pattern is more spread out for nearer objects and is denser for farther objects. The depth data thus forms the 3D mapping data of the space captured by the 3D depth tracker 120.

In some embodiments, the 3D depth tracker 120 includes an RGB (red green blue)
30 or colour sensor integrated with the depth sensor to form an RGB-D camera. The 3D depth tracker 120 with the RGB-D camera may be referred to as an RGB-D tracker 120 and the first tracks generated by the RGB-D tracker 120 may be referred to as RGB-D tracks. The RGB-D tracker 120 is thus configured to detect and track human

subjects based on RGB-D data captured from the human subjects, specifically at their torsos. In one embodiment, the RGB-D tracker 120 has an effective range of approximately 0.5 to 4 m, a horizontal field of view of approximately 90°, and a sampling time of approximately 30 Hz.

5

The 2D laser tracker 140 is configured to perform at least the steps 208, 210, and 212 of the method 200. The 2D laser tracker 140 includes a planar laser scanner or sensor which is available on autonomous robots due to its reliability and accuracy for mapping and navigation tasks. The 2D laser tracker 140 is thus suitable for performing detection and ranging tasks on surfaces. In some embodiments, the 2D laser tracker 140 includes a 2D LiDAR (Light Detection and Ranging) scanner or sensor. The 2D depth tracker 140 may be referred to as a 2D LiDAR tracker 140 and the second tracks determined by the 2D LiDAR tracker 140 may be referred to as 2D LiDAR tracks. Without being limiting, LiDAR may be defined as a surveying method that measures distance to an object by illuminating the object with laser light and measuring the reflected light. Differences in laser return times and wavelengths can then be used to form the 2D mapping data of the space captured by the 2D LiDAR tracker 140.

The 2D LiDAR tracker 140 is thus configured to detect and track human subjects based on 2D LiDAR data captured from the human subjects. Specifically, the 2D LiDAR data is captured from the lower extremities, e.g. legs, of the human subjects. In one embodiment, the 2D LiDAR tracker 140 has an effective range of approximately 0.05 to 10 m, a horizontal field of view of approximately 180°, and a sampling time of approximately 40 Hz. In another embodiment, the 2D LiDAR tracker 140 has a wider field of view, such as up to 270°.

The 2D LiDAR tracker 140 is able to detect human subjects in one plane and the relatively wider field of view allows the 2D LiDAR tracker 140 to track the human subjects with lower risk of losing them out of the field of view. However, the 2D mapping data captured by the 2D LiDAR tracker 140 may be too sparse to be conclusive for tracking the human subjects. The RGB-D tracker 120 is able to generate much denser 3D mapping data, enabling for more features of the human

subjects to be extracted to facilitate tracking. However, the RGB-D tracker 120 has a relatively narrower field of view and a shorter detection range, increasing the risk of losing the human subjects being tracked from the field of view.

- 5 By associating the RGB-D tracks and 2D LiDAR tracks together, the different tracking characteristics of the RGB-D tracker 120 and 2D LiDAR tracker 140 complement each other for collaborative tracking of the human subjects. The disadvantageous characteristics of either tracker can be mitigated by the other. Thus, the device 100 and method 200 can achieve better tracking results in terms of
- 10 reliability and accuracy, compared to tracking the human subjects using either the RGB-D tracker 120 or 2D LiDAR tracker 140 alone.

In many embodiments, the steps 202, 204, and 206 performed by the 3D depth tracker 120, such as the RGB-D tracker 120, are further described in a process 300

15 with reference to Figure 3. The RGB-D tracker 120 is configured to capture 3D mapping data in the form of 3D point cloud data which includes depth data 122 and optionally RGB data 124. In a step 302 of the process 300, the RGB-D tracker 120 constructs 3D representations of the space from the 3D mapping data, wherein the 3D representations are stixel representations or models. The 3D mapping data are

20 captured continually over a time period and at discrete times within the time period, and each stixel representation is constructed for each discrete time. Each stixel representation or model uses vertically oriented rectangles or rectangular sticks known as stixels as elements to form the stixel representations. The stixel representations are constructed to model the space more compactly as the raw 3D

25 mapping data can be significantly reduced to a number of stixels while still accurately representing relevant scene structures in the space, particularly the human subjects and any other objects or obstacles. By representing the space more compactly using stixel representations, further data processing to track the human subjects can be performed quicker and more efficiently, as compared to having to process the raw

30 3D mapping data in a brute-force manner.

Figure 4A illustrates a space or scene captured in a frame 400 at a discrete time by the RGB-D tracker 120. The space contains a number of proposals 410 of the

human subjects, each proposal bounded by a 2D bounding box 420. Figure 4A illustrates five proposals 410a-e with respective bounding boxes 420a-e, and the numbers indicated above each proposal 410a-e is the height (metres) of the respective bounding box 420a-e. In a step 304, the RGB-D tracker 120 detects the proposals 410 of the human subjects from the stixel representations based on prior human data. The proposals 410 represent candidates of possible human subjects that are detected based on the prior human data. The prior human data includes, but is not limited to, data on the human physical shape, such as height dimensions. For example, the RGB-D tracker 120 detects stixels that possibly represents the human subjects based on heights of the stixels. The proposals 410 are iteratively detected by the RGB-D tracker 120 in successive frames 400 at the respective discrete times. Each frame 400 may have initial conditions determined from the previous frame 400, thus allowing for real-time adaptation to small changes in the space or scene, such as slope / terrain variations in the ground and/or camera oscillations of the RGB-D tracker 120 which may be caused by movements of the device 100. The RGB-D tracks are subsequently generated from the iterative detections of the proposals 410.

In one embodiment, the stixel representations are constructed and the proposals 410 are detected through a process 430 as shown in Figure 4B. In a step 432 of the process 430, the RGB-D tracker 120 captures the 3D mapping data or point cloud data for constructing the stixel representations. In a step 434, the RGB-D tracker 120 detects the ground from the point cloud data using various algorithms or models, such as the RANSAC (random sample consensus) method or variants thereof. In a step 436, the RGB-D tracker 120 generates a height map of the points in the point cloud data relative to the detected ground. In a step 438, the RGB-D tracker 120 constructs the stixel representations based on the coordinates of each point relative to the ground. In a step 440, the RGB-D tracker 120 divides the stixel representations into 2D grids and assigns each point to one grid. In a step 442, the RGB-D tracker 120 selects the grids as the proposals 410 based on local density and local maximum height. A predefined parameter range is determined based on the prior human data for bounding the maximum height. For example, the predefined parameter range is 1.2 to 2 m which is the height range of most adults. In the step 442, the grids with the local maximum height bounded within the predefined

parameter range are selected as the proposals 410. In a step 444, the RGB-D tracker 120 selects the points in the selected grids with the local maximum height as anchor points of the proposals 410, as represented by Expression 1 below. In a step 446, the RGB-D tracker 120 computes the 2D bounding boxes 420 of the proposals 410 based on a pinhole camera model. Specifically, the RGB-D tracker 120 is calibrated and the distances of the anchor points from the RGB-D tracker 120 are factored in the computation of the 2D bounding boxes 420. The pinhole camera model is represented by Expressions 2 and 3 below.

10 [Expression 1]

$$(x_{a_i}, y_{a_i}), i = 0 \dots n$$

[Expression 2]

$$H_{proposal} = f_{camera} * \left(\frac{H_{object}}{D_{object}} \right)$$

15

[Expression 3]

$$W_{proposal} = 0.41 * H_{proposal}$$

In Expression 1, (x_{a_i}, y_{a_i}) represents the coordinates of an anchor point of a proposal 410, i represents the ordinal number of the anchor points, and n represents the number of proposals 410. In Expressions 2 and 3, $H_{proposal}$ and $W_{proposal}$ respectively represent the height and width (in pixels) of the proposals 410, f_{camera} represents the focal length (in pixels) of the RGB-D tracker 120, and H_{object} and D_{object} respectively represent the height and distance (in metres) of the objects from the RGB-D tracker 120. The proposals 410 of the human subjects are computed based on the anchor points and corresponding $H_{proposal}$ and $W_{proposal}$.

In another embodiment, the stixel representations are constructed and the proposals 410 are detected through a process 450 as shown in Figure 4C. In a step 452 of the process 450, the RGB-D tracker 120 captures the 3D mapping data or point cloud data for constructing the stixel representations. In a step 454, the RGB-D tracker 120 inputs the ground data which has been pre-calculated using various algorithms or

30

models, such as the RANSAC method or variants thereof. The ground can be calculated offline prior to tracking for many applications of the RGB-D tracker 120. For example, the RGB-D tracker 120 is commonly used on level ground for tracking and since the ground is already known to be level, it can be calculated offline prior to tracking. In a step 456, the RGB-D tracker 120 generates a height map of the points in the point cloud data relative to the ground. In a step 458, the RGB-D tracker 120 divides the ground into grids of predefined cell sizes. For example, each grid has a cell size of 0.1 x 0.1 m which is near to the human head size, so that human subjects can be separated from obstacles such as walls and the human subjects near the walls can be detected. In a step 460, the RGB-D tracker 120 constructs the stixel representations by assigning each point to one grid. In a step 462, the RGB-D tracker 120 records the grid density and local maximum height for each grid. In a step 464, the RGB-D tracker 120 uses non-maximum suppression based on the Ω shape formed by the human head-shoulder structure to detect anchor points over a grid aggregate of 6 x 6 grids. For example, if the grid cell size is 0.1 x 0.1 m, then the anchor points having local maximum heights are detected over a grid aggregate size of 0.6 x 0.6 m, which is about human size. In a step 466, the RGB-D tracker 120 eliminates grid aggregates based on a predefined grid density threshold. The step 466 can eliminate grid aggregates with random noise or very thin objects such as tabletop surfaces. In a step 468, the RGB-D tracker 120 generates a 3D bounding box for each remaining grid aggregate and centres the 3D bounding box at the anchor point of the grid aggregate. Following the examples above, each 3D bounding box has dimensions 0.6 x 0.6 x H m, where H represents the height of the respective anchor point. In a step 470, the RGB-D tracker 120 computes the 2D bounding boxes 420 of the proposals 410 from the 3D bounding boxes. Specifically, the 2D bounding boxes 420 are generated by calibrating the RGB-D tracker 120 and projecting the 3D bounding boxes.

Therefore, by using the stixel representations constructed from the 3D mapping data, the detection of proposals 410 of human subjects is quicker since non-human objects can be eliminated. For example, walls are eliminated because they are higher than the people, ground is eliminated as it is much lower than people, and thin objects such as standing fans are eliminated because they are much thinner than

people. Similarly, objects that are significantly higher, lower, smaller, and/or larger than average dimensions of people can be eliminated, thus making detection of the proposals 410 more efficient and accurate. With the assistance of calibration of the RGB-D tracker 120, the RGB-D tracker 120 is able to determine the position of the proposals 410 as well as the scale of the human subjects in the proposals 410.

With reference to Figure 3, the process 300 includes a step 306 of extracting feature data from the proposals 410. As the RGB-D tracker 120 is able to generate denser 3D mapping data, more feature data, such as including the RGB data 124 or colour features, can be extracted. The extracted feature data may be used to improve tracking of the human subjects based on their RGB-D tracks. Specifically, the extracted feature data of the human subjects within the proposals 410 are compared across successive frames 400 and proposals 410 to achieve continuous tracking of the human subjects. In other words, if the feature data of a human subject in one frame 400 is similar to another human subject in a preceding or succeeding frame 400, the human subject is likely to be identified the same one who is being tracked.

The process 300 includes a step 308 of verifying the proposals 410 that represent or contain the human subjects. Said verifying may be performed by a trained image classifier, such as a support vector machine (SVM) classifier or one based on a convolutional neural network (CNN). For example, a conventional CNN-based classifier can be used to verify if a proposal 410 represents a human subject, while a Siamese CNN classifier can be used to verify if the human subject in the verified proposal 410 is the same as the one being tracked. It will be appreciated that the image classifiers are trained in various ways known to the skilled person. By performing verification of the proposals 410, tracking errors that are accumulated over long term tracking of human subjects can be reduced or eliminated, and the number of proposals 410 is reduced so that the computation resource requirements and cost are low.

In one example, a human subject is identified as a target and is being tracked by the RGB-D tracker 120. After a short period while tracking the target, the target may go missing from the space or scene and the next frame 400 captured by the RGB-D

tracker 120 would not contain the target. The target may go missing for various reasons, such as the target turning around a corner or an obstacle blocks the target in the field of view of the RGB-D tracker 120. The RGB-D tracker 120 may be configured to re-identify the target from the human subjects being tracked. Said re-identifying is based on the feature data of the target and may be performed using the Siamese CNN classifier. The Siamese CNN classifier compares a proposal 410 of the target against other proposals 410 of other human subjects based on their feature data. The target would be re-identified if the comparison results or scores satisfy a predefined threshold. The re-identification capability after a short period when the target goes missing is thus useful for continuous tracking of the target.

In some embodiments, in the step 306, the feature data of the human subjects are extracted by applying a segmentation mask to the proposals 410. Specifically, after the proposals 410 are detected, the proposals 410 are segmented based on the 3D coordinates of points within the proposals 410. The proposals 410 are segmented with the constraint or segmentation mask as shown in Expression 4 below.

[Expression 4]

$$Mask(i, j) = \begin{cases} 0, & \sqrt{(x_i - x_{a_i})^2 + (y_i - y_{a_i})^2} > 0.3m \\ 1, & \sqrt{(x_i - x_{a_i})^2 + (y_i - y_{a_i})^2} < 0.3m \end{cases}$$

In Expression 4, (x_i, y_i) represents the coordinates of a point within the proposal 410, and (x_{a_i}, y_{a_i}) represents the coordinates of an anchor point within the proposal 410. Since all the height of the points or pixels within the proposals 410 is known, all the points can be aligned based on its height. The detected proposals 410 are aligned based on the height from the ground of each point or pixel. The aligned proposals 410 are segmented to segment the human subjects within the proposals 410 from the background. These steps achieve pixel-level height-aligned proposals or templates 480 with segmented human subjects. Figure 4D illustrates the aligned proposals 480a-e corresponding to the original proposals 410a-e.

The aligned proposals 480 may be used to handle or resolve occlusions of partial views of the human subjects. Presence of occlusions in the captured frames 400 and proposals 410 can cause problems in the tracking of human subjects as a human subject may be temporarily blocked by another object or obstacle. For example, a pedestrian being tracked may be blocked by a vehicle or lamppost or other building structures. With the aligned proposals 480, the occlusions can be estimated. These occlusions may include, but are not limited to, occluded body parts of the human subjects, such as legs, head, torso, or parts thereof, etc. Advantageously, missing and/or occluded parts of the human subjects can be uncovered and added to the feature data of the human subjects, thereby improving feature comparisons across successive frames 400. Moreover, verification of the proposals 410 in the step 306 can be improved based on the aligned proposals 480 with the scaled and segmented human subjects, which enable better feature data to be extracted for comparisons. Thus, when a target is being tracked by the RGB-D tracker 120, the RGB-D tracker 120 is able to lock onto the target even under occlusions or partial views so that the target is less likely to go missing as the target can be re-identified in subsequent frames 400.

With reference to Figure 3, the process 300 includes a step 310 of detecting the human subjects from the verified proposals 410. If a proposal 410 cannot be verified to represent a human subject, the proposal 410 may be classified as a partial proposal. Partial proposals may occur if there are severe occlusions present in the proposals 410 such that human subjects cannot be accurately detected. The process 300 includes a step 312 of generating the RGB-D tracks for the detected human subjects from the verified proposals 410.

In an exemplary case of tracking a specific target, the target can be locked and more accurately tracked if the feature data of the target can be compared and matched. But if the feature data cannot be matched, this would probably mean the target is lost and needs to be re-identified. When the target is lost, the preceding steps are re-initiated and the RGB-D tracker 120 searches for nearby proposals 410 with the last observed feature data of the target to try and generate the RGB-D track for the target. It will be appreciated that other types of feature data may be considered, such as

feature data not associated with the human subjects or targets. These other types of feature data may provide information on the surroundings of the human subjects, such as obstacles or other objects. As the RGB-D tracker 120 has a relatively narrower field of view and a shorter detection range, there is greater risk of losing the target. The process 300 thus includes a step 314 of fusing the RGB-D tracking data with that from the 2D LiDAR tracker 140.

In many embodiments, the steps 208, 210, and 212 performed by the 2D laser tracker 140, such as the 2D LiDAR tracker 140, are further described in a process 500 with reference to Figure 5. The 2D LiDAR tracker 140 is configured to capture 2D mapping data in the form of 2D point data or LiDAR data from the lower extremities, e.g. legs, of the human subjects in the space. 2D representations of the space are then constructed from the 2D LiDAR data. The 2D mapping data are captured continually over a time period and at discrete times within the time period, and each 2D representation is constructed for each discrete time. Notably, the time period and discrete times are the same as those for the RGB-D tracker 120 to enable collaborative tracking of the human subjects using both trackers.

The process 500 can be broadly divided into 3 stages – a first stage of identifying individual legs of the human subjects, a second stage of tracking the individual legs, and a third stage of tracking the human subjects based on data from the tracked legs. The multistage process 500 allows the detection of individual legs and the tracking data of the individual legs to be refined into a more meaningful hypothesis of people positions with respect to the 2D LiDAR tracker 140.

The process 500 includes a step 502 of identifying the individual legs. Although each human subject has two legs and both can be tracked, occlusions can occur that prevent observation of both legs simultaneously at all times. For example, one leg may be blocking the line of sight of the other leg, or one leg may frequently occlude the other when the person is walking. As such, the human subjects are more likely to be tracked by observations of individual legs. To identify the individual legs, the 2D LiDAR tracker 140 segments each 2D representation into clusters. The clusters are candidates to be identified if they match certain characteristics or features to be

categorized as legs. A leg confidence score is then computed for each cluster (represented as z) from a number of features of the observed cluster. In some embodiments, the cluster features include the mean of the inscribed angle variation of the points within the cluster and derived by geometric fitting of a circle in the cluster. The mean normalized error of the inscribed angle variation is represented as I_N in Expression 5 below. I_c represents the calculated mean of the cluster c , and $I_d(z)$ represents the benchmark reference mean data taken at distance z . The cluster features further include the normalized error of the standard deviation of the cluster's IAV (represented as S_N in Expression 6 below). S_c represents the calculated standard deviation of the cluster c , and $S_d(z)$ represents the benchmark reference standard deviation taken at distance z . The normalized error of the number of points within the cluster (represented as P_N in Expression 7 below). P_c represents the number of points in the cluster c , and $P_d(z)$ represents the benchmark reference number of points data taken at distance z . The leg confidence score is computed from I_N , S_N , and P_N (represented as Leg_c in Expression 8 below) to form a final confidence value on whether the cluster comprises a leg. K_I , K_S , and K_P represent weighting constants. It will be appreciated that the cluster features may include any of those mentioned above, as well as others such as linearity, circularity, width, and radius of the legs, and the like.

[Expression 5]

$$I_N = \frac{(|I_c - I_d(z)|)}{I_d(z)}$$

[Expression 6]

$$S_N = \frac{(|S_c - S_d(z)|)}{S_d(z)}$$

[Expression 7]

$$P_N = \frac{(|P_c - P_d(z)|)}{P_d(z)}$$

[Expression 8]

$$Leg_c = \max(1 - K_I I_N - K_S S_N - K_P P_N, 0)$$

The process 500 further includes a step 504 of detecting the individual legs. The 2D LiDAR tracker 140 detects the clusters that have legs of the human subjects based on computed leg confidence scores of the clusters. Specifically, the clusters that have the final confidence values above a predefined threshold value will be detected for tracking the legs in these clusters. The 2D LiDAR tracker 140 then generates the 2D LiDAR tracks from the detected clusters, as further described below. In some cases, a cluster may have two or more legs that are very close to or occluding each other, thus forming a larger cluster. The larger cluster could still be detected and the individual legs separated into two smaller clusters if the local minima of the clusters are not too close to the centre of each cluster.

The 2D LiDAR tracker 140 is preferably mounted on the device 100 at a certain height from the ground at which it can detect people's legs, such as 30 cm from the ground. At this height, the 2D LiDAR tracker 140 is low enough that the torso does not present interference in the 2D mapping data, and is high enough that the legs do not move so fast that they cannot be accurately captured. The 2D LiDAR tracker 140 emits laser light to a plane at the appropriate height and measures the reflected light as returned scan points based on distance measurements taken from the plane. The clusters are formed by the scan points based on a predefined distance threshold, such that any points within the threshold are grouped together as a cluster. The threshold is defined to be small enough to separate a person's two legs into two distinct clusters and that each person does not belong to more than two clusters. In some cases, to mitigate noise in the captured mapping data, clusters containing less than three scan points may be discarded in low-noise environments and clusters containing less than five scan points may be discarded in high-noise environments.

Accordingly, in the step 504, the 2D LiDAR tracker 140 detects the individual legs based on the detected clusters and associated leg confidence scores. The process 500 further includes a step 506 of generating 2D LiDAR tracks for the legs. The leg tracks are generated using a set of predictive filters, such as Kalman filters. Briefly, a Kalman filter is a set of mathematical equations that can be used to determine the future location of an object. By using a series of measurements made over time, the

Kalman filter provides a means to estimating past, present, and future states. The Kalman filter may be seen as a Bayers filter under suitable conditions, as will be readily known to the skilled person. It will be appreciated that calculations related to use of the Kalman filter will be readily known to the skilled person, and are not provided herein for purpose of brevity.

A first Kalman filter is used to estimate a first Kalman filter track for the legs (leg tracks) in the detected clusters. The leg Kalman filter uses a constant velocity motion model with a pseudo-velocity measurement during the Kalman filter update steps. At discrete time k , the leg Kalman filter maintains a set of leg tracks ${}^L\mathbf{X}_k$ represented by Expression 9 below, where N represents the number of legs tracked at each discrete time k . In order to initialize the leg tracks, the human subjects may need to remain stationary for the 2D LiDAR tracker 140 to lock on. Each leg track has a state estimate (represented by Expression 10 below) of a leg's position and velocity in 2D Cartesian coordinates. New leg tracks are generated with zero velocity, and existing leg tracks are updated using the constant velocity motion model. During the update steps of the leg Kalman filter, the leg tracks are processed using a leg observation model represented by Expression 11 below.

[Expression 9]

$${}^L\mathbf{X}_k = \{ {}^LX_k^1, {}^LX_k^2, \dots, {}^LX_k^N \}$$

[Expression 10]

$${}^LX_k^j = [x \ y \ \dot{x} \ \dot{y}]^T$$

[Expression 11]

$${}^LZ_k = {}^LH {}^LX_k + v_k$$

The leg observation model includes position and velocity observations with observation noise, v_k , which is assumed to be Gaussian white noise governed by a covariance matrix, ${}^L\mathbf{R}$. The leg Kalman filter may be fine-tuned to compensate the observation noise. The pseudo velocity measurement is determined from estimation

of the difference from the current state (after the update step) and the previous state, normalized by the time step. Each leg track maintains a notion of confidence as a measure of validity of the tracked leg. The confidence is updated based on the interpretations from measurements and are described in a number of possible cases below.

In a first case, there is a leg track associated with a leg observation. The outcome in this first case is that the corresponding people Kalman filter state and the leg track confidence are updated according to Expression 12 below. ${}^Lc_k^j$ represents the confidence of the leg track, ${}^Ld_k^j$ represents the confidence of the leg observation associated with the leg track, and α represents a coefficient parameter that is tunable.

[Expression 12]

$${}^Lc_k^j = \alpha {}^Lc_{k-1}^j + (1 - \alpha) {}^Ld_k^j$$

In a second case, there is a leg track that cannot be associated with any leg observation. The outcome in this second case is that the leg Kalman filter state update step is skipped, but the leg track is propagated using the leg Kalman filter predict step. However, the leg track confidence is degraded according to Expression 13 below.

[Expression 13]

$${}^Lc_k^j = \alpha {}^Lc_{k-1}^j$$

In a third case, there is a leg observation that cannot be associated with any leg track. The outcome in this third case is that a new leg track is generated with zero velocity and zero confidence.

The process 500 further includes a step 508 of processing the leg tracks for tracking the human subjects associated with the tracked legs. Specifically, each leg track is associated with a person and the respective person is tracked using data from the leg track and associated leg observation. The process 500 further includes a step

510 of generating 2D LiDAR tracks for the human subjects. The people tracks are similarly generated using a set of predictive filters, specifically a second Kalman filter to estimate a second Kalman filter track for each human subject (people tracks) associated with the tracked legs.

5

The people Kalman filter uses a constant acceleration motion model which is similar to the constant velocity motion model but with an additional acceleration term. One reason for using the constant acceleration motion model for the people tracking is that people have walking or movement patterns that accelerates and decelerates periodically. At discrete time k , the people Kalman filter maintains a set of people tracks ${}^P\mathbf{X}_k$ represented by Expression 14 below, where N represents the number of people tracked at each discrete time k . In order to initialize the people tracks, the human subjects may need to remain stationary for the 2D LiDAR tracker 140 to lock on. Each people track has a state estimate (represented by Expression 15 below) of a person's position, velocity, and acceleration in 2D Cartesian coordinates. New people tracks are generated with zero velocity and zero acceleration, and existing people tracks are updated using the constant acceleration motion model. During the update steps of the people Kalman filter, the people tracks are processed using a people observation model represented by Expression 16 below.

20

[Expression 14]

$${}^P\mathbf{X}_k = \{ {}^P\mathbf{X}_k^1, {}^P\mathbf{X}_k^2, \dots, {}^P\mathbf{X}_k^N \}$$

[Expression 15]

25 ${}^P\mathbf{X}_k^j = [x \ y \ \dot{x} \ \dot{y} \ \ddot{x} \ \ddot{y}]^T$

[Expression 16]

$${}^P\mathbf{Z}_k = {}^P\mathbf{H}^P + v_k$$

30 The people observation model includes position, velocity, and acceleration observations with observation noise, v_k , which is assumed to be Gaussian white noise governed by a covariance matrix, ${}^P\mathbf{R}$. The people Kalman filter may be fine-

tuned to compensate the observation noise. The position observations are taken from the associated leg tracks described above. Each people track maintains a notion of confidence as a measure of validity of the tracked person. The confidence is updated based on the interpretations from measurements together with that of the tracked legs and are described in a number of possible cases below.

In a first case, there is a people track associated with an observation of two tracked legs. The outcome in this first case is that the corresponding people Kalman filter state and the people track confidence are updated according to Expression 17 below.

$^P c_k^j$ represents the confidence of the people track, $^P d_k^j$ represents the confidence of the tracked person associated with the people track, and β represents a coefficient parameter that is tunable.

[Expression 17]

$$^P c_k^j = \beta ^P c_{k-1}^j + (1 - \beta) ^P d_k^j$$

In a second case, there is a people track that cannot be associated with any tracked leg. The outcome in this second case is that the people Kalman filter state update step is skipped, but the people track is propagated using the people Kalman filter predict step. However, the people track confidence is degraded according to Expression 18 below.

[Expression 18]

$$^P c_k^j = \beta ^P c_{k-1}^j$$

In a third case, there is a people track that is associated with an observation of one tracked leg. The outcome in this third case is that if the one-leg observation is determined to belong to said people track, the corresponding people Kalman filter state and the people track confidence are updated using the properties of the one-leg observation. However, if the one-leg observation cannot be determined to belong to said people track, i.e. the one-leg observation is ambiguous and may belong to

other people track(s), the corresponding people Kalman filter state is not updated and the predict step is skipped.

In a fourth case, there an observation of two legs that cannot be associated with any people track. The outcome in this fourth case is that a new people track is generated with zero velocity, zero acceleration, and zero confidence.

The process 500 further includes a step 512 of processing the people tracks for tracking the human subjects. The process 500 further includes a step 514 of generating the 2D LiDAR tracks for the human subjects from the leg tracks and people tracks. By using the leg and people Kalman filters in a cascading sequence – the leg Kalman filter for tracking individual legs and the people Kalman filter for tracking people – tracking of people can be more effective and robust with lower risk of losing a target.

In the steps 214 and 216 of the method 200, the track fusion module 160 associates the RGB-D tracks and 2D LiDAR tracks for the respective human subjects. The track fusion module 160 then collaboratively tracks each human subject based on the respective associated RGB-D tracks and 2D LiDAR tracks.

In some embodiments, the RGB-D tracks and 2D LiDAR tracks are associated using a Global Nearest Neighbour (GNN) data association method. The GNN data association method presents an uncertainty problem of matching new tracking data to tracks from the previous time $k-1$ (represented as tracks X_{k-1}) to produce updated tracks for the current time k (represented as tracks X_k). This data association problem can be solved by various algorithms, such as the Munkres assignment algorithm. The Munkres algorithm finds the best association that minimizes the total cost according to Expression 19 below, where d represents the Mahalanobis distance between an RGB-D track i and a 2D LiDAR track j .

[Expression 19]

$$\sum D_{ij} = \sum (d_{ij})^2$$

Said associating of the RGB-D tracks and 2D LiDAR tracks may include computing a set of tracking confidence scores from the respective RGB-D tracks and 2D LiDAR tracks for each human subject. The tracking confidence scores include an RGB-D confidence score (represented as C_{RGB-D}) for each RGB-D track, a 2D LiDAR confidence score (represented as $C_{2D-LiDAR}$) for each 2D LiDAR track, and a combined or final confidence score represented as C_{final}) for a combination of the RGB-D and 2D LiDAR tracks. The final confidence score represents a probabilistic aggregation computed from tracking data from the RGB-D tracker 120 and 2D LiDAR tracker 140.

The tracking confidence scores are computed according to Expressions 20 to 23 below. W_1 represents a weight parameter derived from the distance between the RGB-D and 2D LiDAR tracks at the current time, W_2 represents a weight parameter derived from the distance between the individual 2D LiDAR tracks at the current time and previous time, and W_3 represents a weight parameter derived from the distance between the individual RGB-D tracks at the current time and previous time. L_t represents the distance of each track at time t . The weight parameters are defined based on the relative distances of the tracks, each weight parameter should be between 0 and 1 inclusive. If the relative distance of a track becomes too large or too small, the corresponding weight parameter is expected to decrease or increase respectively. It will be appreciated that the Munkres algorithm may use another cost metric instead of the Mahalanobis distance.

[Expression 20]

$$C_{final} = \frac{W_1}{2} * (W_2 * C_{2D-LiDAR} + W_3 * C_{RGB-D})$$

[Expression 21]

$$W_1 = \exp \left\{ - \left[\frac{1}{t} * (L_t^{2D-LiDAR} - L_t^{RGB-D}) \right]^6 \right\}$$

[Expression 22]

$$W_2 = \exp \left\{ - \left[\frac{1}{t_2} * (L_t^{2D-LiDAR} - L_{t-1}^{2D-LiDAR}) \right]^6 \right\}$$

[Expression 23]

$$W_3 = \exp \left\{ - \left[\frac{1}{t_2} * (L_t^{RGB-D} - L_{t-1}^{RGB-D}) \right]^6 \right\}$$

5 In an exemplary scenario, if a target human subject is within the fields of view of the RGB-D tracker 120 and 2D LiDAR tracker 140, the associated confidence scores would be high. However, if for example the target moves away from the narrower field of view of the RGB-D tracker 120, the associated confidence score would be low. Nevertheless, the target can still be tracked if he/she is still within the wider field
10 of view of the 2D LiDAR tracker 140, and the associated confidence remains high. Subsequently, the target can still be re-identified by the RGB-D tracker 120 when the target returns to the field of view of the RGB-D tracker 120. Accordingly, such collaborative tracking of the target using a combination of the RGB-D tracker 120 and 2D LiDAR tracker 140 is advantageous as both trackers complement each other
15 to improve the tracking results.

In some embodiments, the device 100 is a security or surveillance device for monitoring movement patterns of the human subjects. In some other embodiments, the device 100 is a robot or robotic device configured to perform the method 200
20 described above for tracking human subjects. Specifically, the robotic device 100 is of the autonomous type that includes the pursuit controller module 180 for pursuit of or following one of the tracked human subjects according to said collaborative tracking. The robotic device 100 may be a service robot, known as the ISERA, configured for providing services to the human subjects.

25

The pursuit controller module 180 may also be known as an object following controller module. The robotic device 100 has an observation or perception area in its vicinity. The observation area is represented by a spatial map 600 as shown in Figure 6A and Figure 6B. The pursuit controller module 180 determines a strategy
30 for the robotic device 100 to track and follow a target person 610 in motion or at rest based on the spatial map 600. The strategy may be implemented in the form of minimizing the distance between the robotic device 100 and the target person 610,

and simultaneously maximizing safety or distance between the robotic device 100 and surrounding obstacles 620.

- 5 The pursuit controller module 180 first determines one of the tracked human subjects as the target person 610. The target person 610 may be determined based on manual user input or predefined conditions. In one example, the pursuit controller module 180 identifies the human subject closest to the robotic device 100 as the target person 610. In another example, the pursuit controller module 180 identifies the human subject who enters the field of view(s) of the RGB-D tracker 120 and/or
- 10 2D LiDAR tracker 140 as he/she may be a possible intruder, especially if the field of view(s) did not have any human subjects initially. The target person 610 may alternatively be selected by a human user or operator, such as by a gesture command or by inputting image data (e.g. photo) of the target person 610.
- 15 The pursuit controller module 180 then controls motion of the robotic device 100 to follow the target person 610 according to said collaborative tracking of the target person 610. The robotic device 100 may have a set of actuation or motion mechanisms for autonomously moving the robotic device 100. The actuation mechanisms may include wheels and/or continuous tracks (also known as tank threads).
- 20 The robotic device 100 may be of the differential drive or differential wheeled type and including a proportional-integral-derivative (PID) controller module for controlling the linear and angular velocities of the robotic device 100. In controlling motion of the robotic device 100, the pursuit controller module 180 and/or PID controller module calculates the linear and angular velocities to move the robotic
- 25 device 100 towards the target person 610. The calculation is described as a process 700 with reference to Figure 7.

The process 700 includes a step 702 of dividing the spatial map 600 into a plurality of discrete spatial zones or buckets 630. For example, as shown in Figure 6A and

30 Figure 6B, the spatial map 600 has a semi-circular form centred on the robotic device 100, and the spatial zones 630 have identical sector forms. The spatial map 600 may be expanded and reduced proportionally to the distance to the target person 610. The process 700 includes a step 704 of calculating an angular

difference (represented as δ_{Bn}) between each spatial zone 630 relative to the target person 610, according to Expression 24 below. θ_{Bn} represents the angular position of a spatial zone 630 relative to the current heading or direction of the robotic device 100, and θ_P represents the angular position of the target person 610 relative to the current direction. The process 700 includes a step 706 of calculating a cost (represented as $Cost_{Bn}$) of each spatial zone 630 according to Expression 25 below. The cost is calculated based on the respective angular difference and a distance (represented as C_{Bn}). The distance C_{Bn} extends from the robotic device 100 to the nearest obstacle 630 along the respective spatial zone 630, or the radial length of the spatial zone 630 if there is no obstacle 630 along the radial length.

[Expression 24]

$$\delta_{Bn} = |\theta_{Bn} - \theta_P|$$

[Expression 25]

$$Cost_{Bn} = \frac{\delta_{Bn} + 1}{C_{Bn}}$$

The cost of each spatial zone 630 is also influenced by the cost of its neighbouring spatial zones 630. By comparing the average cost of each spatial zone 630, the centre point of the spatial zone 630 with the lowest average cost is chosen to be the pursuit heading of the robotic device 100 for following and pursuit of the target person 610.

The process 700 includes a step 708 of calculating an angular velocity (represented as ω_P) of the robotic device 100 to follow the target person 610, according to Expression 26 below. K_ω represents a weighting constant, and θ_{BP} represents the angular difference between the current direction and pursuit direction of the robotic device 100. The process 700 includes a step 710 of calculating a linear velocity (represented as v_P) of the robotic device 100 to follow the target person 610, according to Expression 27 below. K_o represents a weighting constant, C_{BP} represents the forward distance to move along the pursuit direction, and θ_m

represents an angular difference between the front and edge of the chosen spatial zone 630.

[Expression 26]

$$5 \quad \omega_P = K_\omega \times \theta_{BP}$$

[Expression 27]

$$v_P = \frac{c_{BP}}{K_o} \times \frac{\theta_m - \theta_P}{\theta_m} \times \frac{\theta_m - \theta_{BP}}{\theta_m} \times D_P$$

10 D_P represents a parameter that changes the changes the aggressiveness of the pursuit based on the social zone or immediate surroundings of the target person 610. In some embodiments, typical values for D_P are shown in Expression 28 below where R_P represents the radial distance of the target person 610 from the robotic device 100.

15

[Expression 28]

$$D_P = \begin{cases} 1 & \text{amp; if } R_P > 1.5m \\ 0.6 & \text{amp; if } 1m < R_P < 1.5m \\ 0.2 & \text{amp; if } 0.3m < R_P < 1m \\ 0 & \text{amp; if } R_P < 0.3m \end{cases}$$

Accordingly, the pursuit controller module 180 controls motion of the robotic device 20 100 based on the linear and angular velocities calculated in the process 700. The robotic device 100 is thus configured to track and follow the target person 610.

While following the target person 610, he/she may go missing from the field of view of the RGB-D tracker 120, such as if the target person 610 suddenly disappears into 25 a sharp turn. As described above, as the trail of the target person 610 is still being tracked by the 2D LiDAR tracker 140, the robotic device 100 can still follow the last known position of the target person 610 and the RGB-D tracker 120 is able to re-detect the target person 610 when he/she returns to its field of view. A trained image classifier such as a Siamese CNN classifier attempts to match the last observed 30 feature data of the target person 610 to those of other tracked human subjects. If the

feature data matches one of the tracked human subject, then said tracked human subject is re-identified as the target person 610. Conversely, if the feature data does not match any of those of other tracked human subjects, the classifier discards the human subjects at the current time and continues matching for subsequent times.

5 Thus, this re-identification capability of the robotic device 100 allows it to lock onto the target person 610 and continuously track and follow the target person 610 using a combination of the RGB-D tracker 120 and 2D LiDAR tracker 140.

10 There are many potential applications of people following in the various fields including the service and healthcare industries. For example, a human-following assistive robot increases the capacity of a person in transporting goods, carrying luggage, lifting of heavy objects etc. by means of providing for the additional payload. The assistive robot can follow the person to a desired location to unload the heavy objects or load. In another example of a security or surveillance robot in a secured
15 premise, the robot can detect a person, regardless if he/she is a friendly or possible intruder, and approach the person to request for further clarification, such as verification of identity. In another example, an assistive robot in medical / healthcare can help a patient in rehabilitation following an injury by assessing the patient's walking or movement patterns.

20

As described in various embodiments herein, the device 100 and method 200 are able to perform collaborative tracking of people by adopting a combination of two modalities from the 3D depth tracker 120 and 2D laser tracker 140. These trackers are described more specifically as the RGB-D tracker 120 for acquiring colour and
25 depth information and the 2D LiDAR tracker 140 for acquiring 2D LiDAR scans. Information from both scanners are fused together and this combination enables their different tracking characteristics to complement each other for collaborative tracking of people. An advantage is that the accumulation of errors in tracking of people, especially over an extended time period, is reduced or minimized. This error
30 accumulation may occur during detection of proposals 410 by the RGB-D tracker 120, is about two to three orders of magnitude. Thus, the device 100 and method 200 for collaborative tracking of people can achieve better tracking results in terms of reliability and accuracy.

In the foregoing detailed description, embodiments of the present disclosure in relation to a device and method for tracking human subjects are described with reference to the provided figures. The description of the various embodiments herein
5 is not intended to call out or be limited only to specific or particular representations of the present disclosure, but merely to illustrate non-limiting examples of the present disclosure. The present disclosure serves to address at least one of the mentioned problems and issues associated with the prior art. Although only some embodiments of the present disclosure are disclosed herein, it will be apparent to a person having
10 ordinary skill in the art in view of this disclosure that a variety of changes and/or modifications can be made to the disclosed embodiments without departing from the scope of the present disclosure. Therefore, the scope of the disclosure as well as the scope of the following claims is not limited to embodiments described herein.

Claims

1. A device for tracking human subjects, the device comprising:
- a 3D depth tracker comprising an RGB-D camera and configured for:
 - capturing 3D mapping data of a space containing one or more human subjects;
 - constructing 3D stixel representations of the space from the 3D mapping data;
 - detecting proposals of the human subjects from the 3D stixel representations based on prior human data; and
 - generating a first track for each human subject in each 3D stixel representation from the proposals, the first tracks for tracking the respective human subject;
 - a 2D laser tracker having a wider field of view than the 3D depth tracker, the 2D laser tracker configured for:
 - capturing 2D mapping data of the space;
 - constructing 2D representations of the space from the 2D mapping data; and
 - generating a second track for one or both legs of each human subject in each 2D representation, the second tracks for tracking the legs of the respective human subject; and
 - a track fusion module configured for:
 - associating the respective first tracks with the respective second tracks for each human subject; and
 - collaboratively tracking each human subject based on probabilistic aggregation of the respective associated first and second tracks,
- wherein the 3D depth tracker is configured for, during said collaborative tracking of the human subjects, re-identifying a target among the tracked human subjects after the target leaves and returns to the 3D depth tracker's field of view.

2. The device according to claim 1, wherein the 2D laser tracker is further configured for:

segmenting each 2D representation into clusters;

computing a leg confidence score for each cluster; and

5 detecting the clusters that comprise legs of the human subjects based on the computed leg confidence scores of the clusters;

wherein the second tracks are generated from the detected clusters.

3. The device according to claim 1, wherein the device is a robotic device
10 comprising a pursuit controller module configured for:

determining one of the tracked human subjects as the target; and

controlling motion of the robotic device to follow the target according to said collaborative tracking of the target,

15 wherein the 3D depth tracker is configured for re-identifying the target based on feature data of the target for continuous tracking of the target.

4. The device according to claim 1, wherein the 2D laser tracker comprises a 2D LiDAR scanner.

20 5. A method for tracking human subjects, the method comprising:

capturing 3D mapping data of a space using a 3D depth tracker comprising an RGB-D camera, the space containing one or more human subjects;

25 constructing 3D stixel representations of the space from the 3D mapping data;

detecting proposals of the human subjects from the 3D stixel representations based on prior human data;

30 generating a first track for each human subject in each 3D stixel representation from the proposals, the first tracks for tracking the respective human subject;

capturing 2D mapping data of the space using a 2D laser tracker having a wider field of view than the 3D depth tracker;

constructing 2D representations of the space from the 2D mapping data;

generating a second track for one or both legs of each human subject in each 2D representation, the second tracks for tracking the legs of the respective human subject;

associating the respective first tracks with the respective second tracks for each human subject;

collaboratively tracking each human subject based on probabilistic aggregation of the respective associated first and second tracks; and

during said collaborative tracking of the human subjects, using the 3D depth tracker, re-identifying a target among the tracked human subjects after the target leaves and returns to the 3D depth tracker's field of view.

6. The method according to claim 5, further comprising extracting feature data of the human subjects from the proposals.

7. The method according to claim 5 or 6, further comprising verifying, using a trained image classifier, the proposals that represent the human subjects.

8. The method according to claim 7, wherein the trained image classifier is based on a convolutional neural network.

9. The method according to claim 5, further comprising:

segmenting each 2D representation into clusters;

computing a leg confidence score for each cluster; and

detecting the clusters that comprise legs of the human subjects based on the computed leg confidence scores of the clusters;

wherein the second tracks are generated from the detected clusters.

10. The method according to claim 9, wherein the second tracks are generated using a set of predictive filters.

11. The method according to claim 10, wherein the set of predictive filters comprises:

a first Kalman filter for estimating a first Kalman track for the legs in the detected clusters; and

5 a second Kalman filter for estimating a second Kalman track for each human subject associated with the tracked legs,

wherein the second tracks are generated from the respective first and second Kalman tracks.

10 12. The method according to claim 5, wherein the first and second tracks are associated using a Global Nearest Neighbour (GNN) data association method.

13. The method according to claim 5 or 12, wherein said associating comprises computing a set of tracking confidence scores from the respective first and second
15 tracks for each human subject.

14. The method according to claim 5, wherein the method is performed by a robotic device, the method further comprising:

determining one of the tracked human subjects as the target; and

20 controlling motion of the robotic device to follow the target according to said collaborative tracking of the target;

re-identifying, using the 3D depth tracker, the target from the tracked human subjects based on feature data of the target for continuous tracking of the target.

25

15. The method according to claim 5 or 14, wherein said re-identifying uses a Siamese convolutional neural network.