FACTA UNIVERSITATIS
Series: **Automatic Control and Robotics** Vol. 15, N^{o} 1, 2016, pp. 17 - 32

LOW-LEVEL SENSOR FUSION-BASED HUMAN TRACKING FOR MOBILE ROBOT

UDC (((681.58:004.89)+(167.7:63)):621.317)

Danijela Ristić-Durrant, Ge Gao, Adrian Leu

Institute of Automation, University of Bremen, Germany

Abstract. *In this paper, a novel sensor-based human tracking method that enables a mobile robot to follow a human with high robustness and responsiveness is presented. The method is based on low-level sensor data fusion combining depth data from a stereo camera and an infrared 2D laser range finder (LRF) to detect target human in the near surrounding of the robot. After initial position of target human is located by sensor fusion-based human detection, a novel tracking algorithm that combines a laser data-based search window and Kalman filter is used to recursively predict and update estimations of target human position in robot's coordinate system. The use of tracking window contributes to reduction of computational cost by defining region of interest (ROI) enabling so real-time performance. The performance of proposed system was tested in several indoor scenarios. Experimental results show that the proposed human detection algorithm is robust and human tracking algorithm is able to handle fast human movements and keep tracking of target human in various scenarios.*

Key words: *mobile robot, range finders, stereo vision, target tracking, human robot interaction, Kalman filter*

1. INTRODUCTION

Human tracking has been extremely active research area in the computer vision community over the past decade. The importance of this area arises from its numerous applications such as video surveillance, smart vehicle, virtual reality, and the mobile robot vision. The later represents one of the broadest application areas of vision-based human tracking [1]. Since the ability of environment perception is limited if the robot is relying on information from a single vision sensor, integrated information from multiple sensors can significantly improve the robot perception performance.

Based on different levels of prerequisites and prior knowledge regarding geometry of system setup and sensor traits, the implemented human detection and tracking approaches

Received December 09, 2015

Corresponding author: Danijela Ristić-Durrant
Institute of Automation, University of Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Germany
E-mail: ristic@iat.uni-bremen.de

vary in a wide range. Approaches based on sensor data fusion between stereo camera and 2D laser range finder (LRF) are very promising since they are able to make use of both robustness and accuracy features of laser data and affluent information of image data [2].

In most of the methods in literature on using the combination of vision and laser sensors for human tracking, human detection is conducted with vision sensor and laser sensor separately, and then the detection results are fused [3][4]. In other words, more processed high level segmentation and classification results are fused. In contrast, this paper addresses fusing the low-level data, such as depth information from laser and disparity map from stereo vision system. The idea behind is to develop a sensor fusion algorithm, which fuses low-level sensor data since those sensor data are more reliable and robust as they are less interpreted by various algorithms. Also, it leads to a fast algorithm for real-time tracking application. The real-time aspect of vision-based human detection, which represents a part of novel approach presented in this paper, has been validated in [5]. In this paper performance evaluation of completed system including low-level fusion of vision and laser sensor data is presented.

Beside the novel approach based on low-level sensor depth data fusion for human detection presented in this paper, the contribution of this paper is also a novel approach based on first derivative of laser depth instead of using traditional laser data clustering methods which are based on Euclidean distance between two consecutive laser points such as dynamic distance threshold [6] and jump distance clustering [7]. This approach is presented in Section 4.2. The rest of the paper is organized as follows. Overview of the system is presented in section 2. Section 3 presents stereovision-based human detection. Section 4 contains the first part of developed algorithm, which is low-level data processing for laser and image sensors and fusion human detection. The second part of the algorithm regarding human tracking is presented in section 5. Section 6 provides the performance evaluation of presented fusion-based human tracking in different indoor scenarios.

2. System Overview

Figure 1 shows hardware layout and software architecture of the sensor fusion-based robotic human follower system. The mobile robot platform is equipped with a stereo camera on top, and a 2D LRF in lower part of the robot. The location and orientation of stereo camera ensures that the complete figure of a target human is included inside the camera Field of View (FOV), while the human is approximately 2 meters in front of the robot. The vertical distance from LRF to the floor is approximately 20 cm, which is at the similar height of human ankle-calf area. In the software architecture, Fusion System consists of three main modules, which are *Vision Human*, *Fusion Human* and *Laser Tracker*.

Vision Human module relies on stereo camera as its data source and conducts a stereovision-based human detection based on the segmentation of disparity map. *Fusion Human* obtains input data from both stereo camera and LRF, and undertakes human detection based on low-level data fusion. By comparing detection results from both modules, the target human is located and the result is fed into a Kalman filter to perform fusion tracking process. *Laser Tracker* equips the system with a fast tracking performance. It is supported with a Kalman filter, and is activated by the first matching human detection result from Vision Human and Fusion Human. Fusion tracking process is highly robust but low efficient with approximately 4 Hz data rate, while laser tracking process is less robust but highly efficient with 20 Hz data rate, so that the overall combined human tracking performance is both highly robust and highly efficient. In order to guarantee a robust human detection result, *Vision Human* detection result is used to verify *Fusion Human* detection result before activation of *Laser Tracker* and to supervise tracking performance. Figure 2 represents detailed block-diagrams of *Vision Human* and *Fusion Human*.

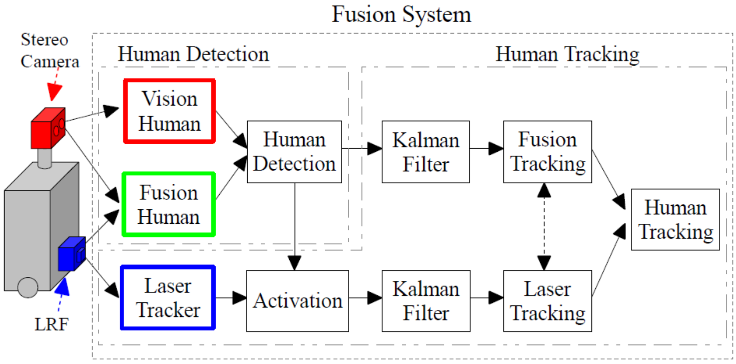


Fig. 1 Layout of mobile robot platform and *Fusion System* overview

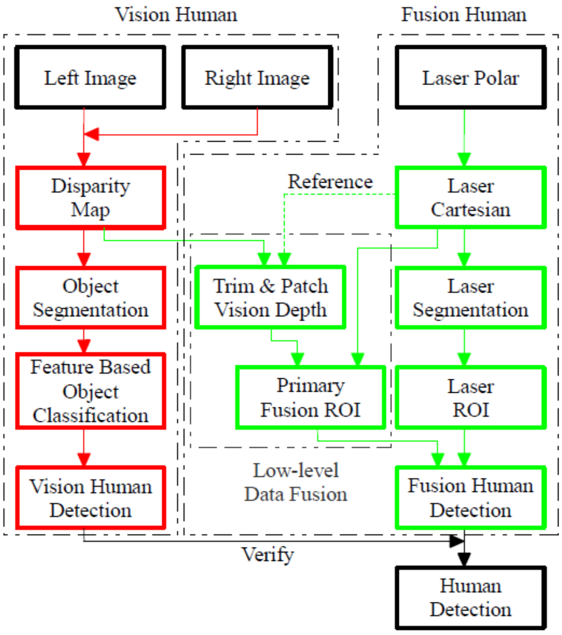


Fig. 2 Detailed diagram of *Vision Human* and *Fusion Human* modules

3. Stereovision-based Human Detection

A block-diagram of the Vision Human module is given in Figure 2, and the individual processing steps are explained in the following sections.

3.1. Object segmentation

Stereo image information is used to aid the definition of regions of different objects, including humans, in camera images. The acquired image pair is used to compute a disparity map using a block-matching algorithm [10]. In principle, the disparity map is found by computing the stereo correspondences between the image points from the left and right stereo images. For a given 3D point $P(X, Y, Z)$, the corresponding points in the left and right image respectively are $p_L(u_L,v_L)$and $p_R(u_R,v_R)$, where u and v are coordinates of an image point in the image coordinate system with the origin in the camera's optical centre. The difference in u coordinates of corresponding points is known as disparity d:

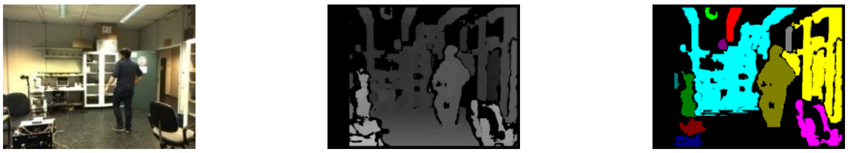
$$d = u_L - u_R \, . \tag{1}$$

The disparity of an image point is inversely proportional to the distance of the original 3D point to the camera coordinate system which is known as depth v_i. The resulting disparity map represents a 2D image in which values of pixels are equal to the disparity (1) and consequently inversely proportional to the depth v_i. The pixel coordinates in a disparity map correspond to the pixel coordinates in the left stereo image as the left stereo image is usually considered as the reference image when building the disparity map.

The resulting disparity map in the proposed system is segmented using a connected pixel labeling-based method. The main idea behind this segmentation method is to group the pixels with the same or very close pixel values as it is assumed that they belong to the same object. Namely, neighbouring pixels in the disparity map belonging to the surface of an object have close disparity values, while on the edges of the object the difference in disparity values between the pixels of the object and of the background is large. These transitions in disparity values are used for the segmentation. The details of the disparity map segmentation method used are given in [11]. The segmentation result in the case of the human tracking robot scenario considered here is shown in Figure 3(c). Differently coloured regions in the image in Figure 3(c) represent different objects, which are at different distances to the robot's camera. As can be seen, the ground has been removed from the segmented image (represented by black colour) in order to avoid merging with other objects, including humans, placed on the ground. The ground plane removal was done by detecting the regions in the lower part of the disparity map whose disparity values gradually change, i.e., whose image gradient in the vertical direction gradually changes. In contrast to ground plane pixels, in the disparity map the regions of object surfaces have almost constant disparity values.

3.2. Feature-based object classification

Once the objects have been segmented in the disparity map, different features describing segmented object regions are calculated. The chosen features have been defined so as to enable distinguishing of humans from other objects in the robot's perceived environment. The used features can be separated into two groups: 2D features and 3D features. The former are calculated from the 2D segmented image, while the later result from 3D object reconstruction.



(a) (b) (c)

Fig. 3 Left stereo image of human walking in front of the robot (a). Disparity map (b). Segmented disparity map (c).

2D features. In order to describe the shape of every segmented region, the three so-called Hu invariant moments [12] are used, as they are proven to be effective shape descriptors. They are calculated over the pixels of each segmented object region according to following formulas:

$$H_1 = \frac{\mu_{20}}{\mu_{00}^{2}} + \frac{\mu_{02}}{\mu_{00}^{2}} . \tag{2}$$

$$H_2 = \left(\frac{\mu_{20}}{\mu_{00}^{2}} - \frac{\mu_{02}}{\mu_{00}^{2}} \right)^2 + \frac{4\mu_{11}}{\mu_{00}^{2}} . \tag{3}$$

$$H_3 = \left(\frac{\mu_{30}}{\mu_{00}^{5/2}} - 3\frac{\mu_{12}}{\mu_{00}^{5/2}} \right)^2 + \left(3\frac{\mu_{21}}{\mu_{00}^{5/2}} - \frac{\mu_{03}}{\mu_{00}^{5/2}} \right)^2 . \tag{4}$$

where μ_{pq} is the central moment defined as:

$$\mu_{pq} = \sum_{u}\sum_{v}(u-\bar{u}_c)^p (v-\bar{v}_c)^q I(u,v), \qquad p,q = 0,1,2,\ldots . \tag{5}$$

In (5) $\bar{u}_c$ and $\bar{v}_c$ are image coordinates of the centre of mass of the segmented object region (average over all image coordinates of segmented pixels in a region). $I(u,v)$ is the intensity level of an image point with coordinates u and v. In the case of a segmented image, $I(u,v)$ is 1 for each pixel segmented as belonging to an object and 0 for each background (not segmented) pixel.

3D features. In order to effectively distinguish humans from other objects in the robot's environment, Hu moments, as descriptors of shapes of segmented object regions in 2D segmented disparity maps, are supported by two 3D objects features: object height and object width. To calculate these two features, first the bounding box of each segmented object region is defined as the smallest rectangle containing the segmented region in the 2D segmented disparity map. Then, the upper left corner and the bottom right corner of each bounding box are identified and their coordinates in the left stereo image are calculated as (u_{LU}, v_{LU}) and (u_{RB}, v_{RB}) respectively. In order to calculate real 3D object width and height, the 3D reconstruction (so-called 2D to 3D mapping [13]) of the bounding box corner points is performed according to:

$$X_{LU} = \frac{b \cdot u_{LU}}{d}, Y_{LU} = \frac{b \cdot v_{LU}}{d}, Z_{LU} = \frac{b \cdot f}{d} . \tag{6}$$

$$X_{RB}=\frac{b\cdot u_{RB}}{d}, Y_{RB}=\frac{b\cdot v_{RB}}{d}, Z_{RB}=\frac{b\cdot f}{d}. \tag{7}$$

where *f* is focal length of the camera (in pixels) and *b* is the stereo camera base line (in meters) representing a line connecting the camera centres of stereo cameras. In (6) and (7), *d* is the disparity of the centre of mass of the segmented object region (u^{-}_{c}, v^{-}_{c}) in the segmented disparity map. With the 3D coordinates (6) and (7) of the corner points, the object height *h* and object width *w* are calculated as:

$$h=Y_{LU}-Y_{RB}, w=X_{RB}-X_{LU}. \tag{8}$$

The defined features are used in the proposed system for the classification of an object as belonging or not belonging to the class “human being”. The used classifier is Back propagation Neural Network with one hidden layer [14]. The training of the classifier, i.e., the neural network parameters adjustment, was done using a training set of 577 feature vectors (H_1, H_2, H_3, h,w) extracted from segmented human regions in disparity maps of stereo image pairs acquired indoors as well as outdoors. The testing of the developed classifier was done using another 423 feature vectors. These test features were obtained by extraction from segmented regions of different objects, including humans, in disparity maps of stereo image pairs acquired indoors as well as outdoors. The obtained classification result from the training set was very good, as indicated by the fact that the classification performance rate was 96%. Misclassification, i.e., the inability to classify an object from a robot’s environment as human, happened in cases of significant human occlusion or segmentation of humans as connected to objects from the environment.

3.3. Vision human detection

The last processing step in the proposed vision system is human detection. Once the human has been successfully classified, his/her 3D position with respect to the coordinate system of the left stereo camera is calculated through the 3D coordinates of the centre of mass of the segmented human region ($\bar{u}_{HC}$, $\bar{v}_{HC}$) in the segmented disparity map:

$$X_{HC}=\frac{b\cdot \bar{u}_{HC}}{d}, Y_{HC}=\frac{b\cdot \bar{v}_{HC}}{d}, Z_{HC}=\frac{b\cdot f}{d}. \tag{9}$$

where coordinate Z_{HC} represents depth.

The coordinates (9) representing coordinates in camera coordinate system can be transformed to coordinates of robot coordinate system (*X*,*Y*,*Z*) knowing the robot’s geometry and using the fact that the robot coordinate system is of the same orientation as camera coordinate system.

In case of object misclassification, Vision Human module does not have an output. As explained in previous section, it happens, for example, in cases of significant human occlusion or segmentation of humans as connected to objects from the environment. This indicates a need for integrating a module to provide an input to mobile robot control, even in the case of human classification failure. In [5] such a module is the modified Kalman filter for prediction and estimation of the 3D position of the human. In this paper, a module that overcomes the problem of not reliable Vision Human Detection is the Fusion Human Detection supported by low-level sensor data fusion as described in following sections.

4. Low-level Sensor Data Fusion

4.1. Fusion points

Low-level data processing and fusion process is only conducted inside the overlapped FOVs of LRF and vision sensor. Figure 4 illustrates the data distribution after extrinsic sensor calibration of stereo camera and LRF [8]. The rectangle denotes the vision FOV while blue and red dots represent the planar FOV of LRF. Fusion data processing is viable only for Fusion Points, which are denoted with red dots. Each Fusion Point ideally should have low-level depth information from both LRF and vision sensors that are processed as described in following sections.

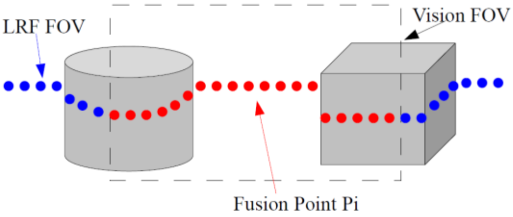


Fig. 4 Fusion points

4.2. Laser data processing and Laser ROI extraction

The raw LRF data is in polar coordinate which includes range r_i and angle θ_i information for each Fusion Point P_i, with r_i being the horizontal distance from a target to LRF origin and θ_i being the angle offset (Figure 5). In order to obtain depth information, conversion between polar and Cartesian coordinates is performed:

$$X_i = r_i \sin\theta_i, Y_i = r_i \cos\theta_i \tag{10}$$

Assuming LRF depth information for P_i is l_i, from (10) it is:

$$l_i = X_i,\ i \in \{0,\dots,n\} \tag{11}$$

where n is the total number of Fusion Points. The vector which contains laser depth information is $\boldsymbol{l}$:

$$\mathbf{l} = [l_0,\dots,l_n] \tag{13}$$

The first derivative of l_i (when $i > 0$) is d_i,

$$d_i = l_i - l_{i-1}, i \in \{1,\dots,n\} \tag{14}$$

and vector **d** contains the first derivatives:

$$\mathbf{d} = [d_1,\dots,d_n], \tag{15}$$

If $d_i > 0$,that is $l_i > l_{i-1}$, P_i is farther from the LRF origin than P_{i-1}. Similarly, if $d_i < 0$, P_i is closer to LRF than P_{i-1}. In order to remove noise and detect significant depth change between two consecutive Fusion Points P_{i-1} and P_i, d_i is filtered with an appropriate threshold t, so that d_i =0 if | d_i| < t.

After obtaining **d**, positive and negative derivatives d_i can be extracted since they are potential indicators of start and end points of objects. In Figure 5, assuming an object is placed in the robot's environment in front of a wall, then first derivative vector **d** contains $d_A < 0$ and $d_{a+1} > 0$ with

$$d_A = l_A - l_{A-1},\, d_{a+1} = l_{a+1} - l_a \tag{16}$$

Points P_A and P_a can be located as potential start and end points of the object. Let **s** be the vector of all Fusion Points P_i with $d_i < 0$ and **e** be the vector of Fusion Points P_i with $d_i > 0$:

$$\mathbf{s} = [P_A, P_A, P_A, \ldots],\ \ \mathbf{e} = [P_a, P_b, P_c, \ldots] \tag{17}$$

For each potential object start point in **s** and the corresponding potential object end point in **e** should be correctly located to form a line segment. For example, in Figure 5 the object line segment is L_{Aa}. The Euclidean distance D_{Aa} between laser point coordinates (X_A, Y_A) and (X_a, Y_a) can be calculated as:

$$D_{Aa} = \sqrt{(X_A - X_a)^2 + (Y_A - Y_a)^2}\ . \tag{18}$$

This paper concerns human detection so that human legs can be considered as object of interest. In order to locate potential human leg segments, constraints can be set according to human body feature, which is the limited width of ankle-calf area. Assuming the upper threshold is D_T , then L_{Aa} belongs to Laser ROI which will be used in future processing, if $D_{Aa} \leq D_T$.

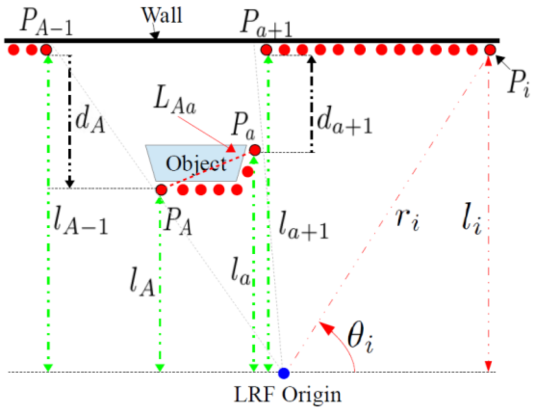


Fig. 5 Laser depth and laser object segments

4.3. Low-level vision data processing

Due to the correspondence problem while obtaining disparity map [9], if the corresponding pixels can not be located in low-texture regions, the depth value inside those regions can not be recovered. In order to improve data quality of vision depth when possible, for each Fusion Point P_i, vision depth v_i is processed using *patching*, which is a process proposed in this paper. Patching retrieves the missing depth value v_i according to its previous neighbour:

$$v_i = \begin{cases} v_{i-1}, \text{ if } \; v_{i-1} \; \text{ exist} \\ 0, \text{ otherwise} \end{cases}, \; i \in \{0, \ldots, n\}, \tag{19}$$

where n is total number of Fusion Points. In case of spurious stereo matches with poor match quality and false vision depth result [8], vision depth *trimming* is proposed in this paper. Trimming means that under certain circumstances, false vision depth can be corrected by using corresponding LRF depth information as the reference. For example, in case of an opaque object, ideally the depth data for a Fusion Point P_i should be $v_i = l_i$. If due to false vision depth, $v_i > l_i$ (Figure 6(a) upper example), the value v_i is replaced with l_i as the reference value. This is because stereo camera can not see through opaque objects. False vision depth can also result in $v_i < l_i$ (Figure 6(a) upper example), however in this case, described correction of v_i using l_i as the reference can not be done. As it is illustrated in lower example in Figure 6(a), the same situation occurs when P_i is laying on an optical transparent object. In this case v_i is correct but laser beam can penetrate optical transparent materials which results in $v_i < l_i$. Proposed trimming process can be expressed as:

$$v_i = \begin{cases} l_i, \text{ if } v_i > l_i \\ v_i, \text{ otherwise} \end{cases}. \tag{20}$$

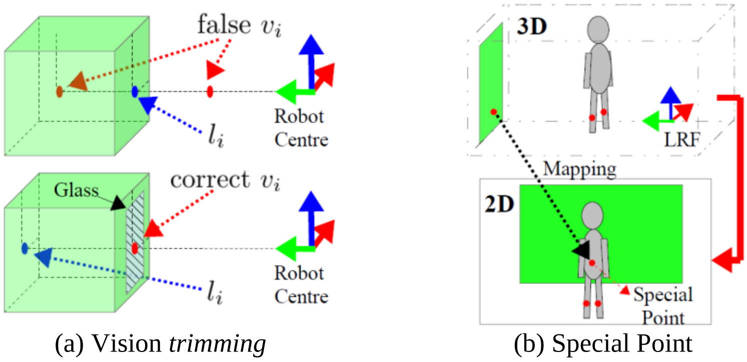
Therefore, after trimming, for each Fusion Point P_i

$$v_i \le l_i, \; i \in \{0, \ldots, n\}. \tag{21}$$

There are two possible causes of $v_i < l_i$. The first is due to remaining false vision depth v_i after trimming, the other is caused by the points which are considered as *Special Fusion Points*, which are projected on the environment's background from LRF in 3D space and mapped on target human in 2D disparity map (Figure 6(b)). A Special Fusion Point has two depth values, laser depth l_i from the background (wall) and vision depth v_i from foreground (target human). Therefore the *Special Fusion Point* is characterized with a depth offset d_{lv}:

$$d_{lv}(i) = l_i - v_i, \; i \in \{0, \ldots, n\} \tag{22}$$

This is caused by width and structure difference between human upper and lower body and the depth offset can be used as a cue for potential human presence.



(a) Vision *trimming* (b) Special Point

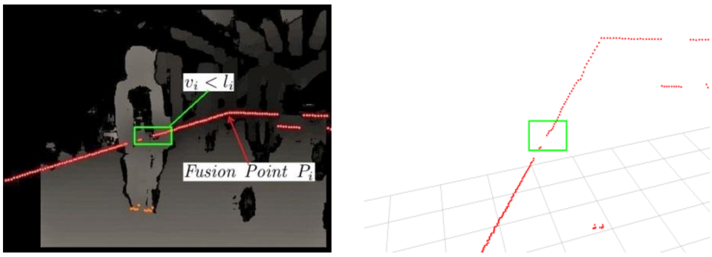
Fig. 6 Vision depth trimming

4.4. Sensor fusion-based human detection

Figure 7 illustrates a scenario with a target human in front of the mobile robot. Figure 7(a) presents laser data projected onto the 2D disparity map with vision depth after patching and trimming where Fusion Points are presented as red dots. Special Fusion Points, which are described in section 4.3, are marked with green rectangle. Corresponding laser points are also marked with green rectangle in Figure 7(b). If depth offset $d_{lv}(i) > 0$, Special Fusion Point P_i belongs to *Primary Fusion ROI*. Due to the remaining false v_i as described in section 4.3, Primary Fusion ROI may contain both potential target human regions, as illustrated in Figure 7(a), and false vision depth regions. Inside Primary Fusion ROI, Fusion Points with consecutive indices are merged into segments F. Vector $\mathbf{F}$ contains the resulted segments:

$$\mathbf{F} = [F_{jk}, F_{lm}, F_{pq} \ldots], \tag{23}$$

where for F_{jk}, j and k are the indices of start and end Fusion Point of this segment. For each segment F_{jk} inside Primary Fusion ROI, if there exists one or more segment(s) L_{ab} from Laser ROI (described in section 4.2) satisfying $b+2 \leq j$ or $k+2 \leq a$, then segment F_{jk} is selected as a correct segment in F_{jk} (Figure 8). Consequently, the segments inside Primary Fusion ROI which do not satisfy above criteria, are filtered out as they are considered as false segments. Final Fusion ROI indicates the detected target human by Fusion Human module.



(a) Disparity map with laser projection (b) Laser data

Fig. 7 *Special Fusion Points* inside green rectangle

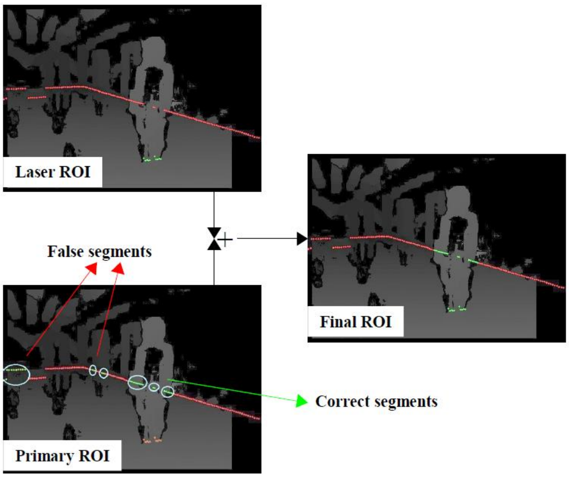


Fig. 8 *Final Fusion ROI* creation

5. Human Tracking

5.1. Laser-based fast human tracking

As shown in Figure 1, two standard Kalman filters [1] are used separately to conduct Fusion Tracking (FT) and Laser Tracking (LT), where FT is used to supervise LT to maintain tracking robustness and LT is used for fast tracking. The human detection result (Final ROI) from above described Fusion Human is verified with human detection from Vision Human described in section 3. This means the FT is initialized if Final ROI is considered as correct detection which is if it is within the human region detected by Vision Human. The (X,Y) Cartesian coordinates of the mean position of Fusion Points inside Final ROI (green dots in Figure 8) are calculated and used as target human position for FT.

Assuming the state of a system at time t is derived from prior state at time $t-1$ [15]

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \mathbf{B}_t \mathbf{u}_t + \mathbf{w}_t \tag{24}$$

where $\mathbf{x}_t$ is system state vector at time t, $\mathbf{u}_t$ is the vector of control inputs, $\mathbf{A}_t$ is the state transition matrix, $\mathbf{B}_t$ is the control input matrix and $\mathbf{w}_t$ is the vector containing process noise. Assuming the model for measurements of the system $\mathbf{z}_t$ [15]

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{v}_t \tag{25}$$

where $\mathbf{z}_t$ is the vector of measurements, $\mathbf{H}_t$ is the transformation matrix which maps state vector into measurement domain and $\mathbf{v}_t$ is the measurement noise vector.

A feedback control with recursive properties is used for predicting which can be described with prediction update and measurement update [16] [15]. From time step $t-1$ to time step t , standard Kalman filter equations for prediction update are:

$$\hat{\mathbf{x}}_{t|t-1} = \mathbf{A}_t \hat{\mathbf{x}}_{t-1|t-1} + \mathbf{B}_t \mathbf{u}_t \tag{26}$$

$$\mathbf{P}_{t|t-1} = \mathbf{A}_t \mathbf{P}_{t-1|t-1} \mathbf{A}_t^T + \mathbf{Q}_t \tag{27}$$

where $\hat{\mathbf{x}}_{t|t-1}$ is the estimation of unknown true system state $\mathbf{x}_t$, covariance matrix $\mathbf{P}_t$ stores the variances and covariances for describing the Gaussian functions and $\mathbf{Q}_t$ is the process noise covariance matrix [16] [15]. The equations for measurement update are:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t (\mathbf{z}_t - \mathbf{H}_t \hat{\mathbf{x}}_{t|t-1}) \tag{28}$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{H}_t \mathbf{P}_{t|t-1}) \tag{29}$$

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^T (\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R}_t)^{-1} \tag{30}$$

where $\mathbf{R}_t$ is the measurement noise covariance, $\mathbf{K}_t$ is the Kalman gain [16] [15]. In Kalman filtering cycle, the prediction update predicts the future state while measurement update corrects the estimated value by the real measurement at present. During tracking process, a single measured value is usually noisy and with the estimated values from Kalman filtering, a more smooth and precise trajectory can be realized.

For Kalman filter in FT, the prediction states include the position and velocity of target human where the position is expressed via two Cartesian coordinates, *X* and *Y* coordinates in robot coordinate system,. Measurement state is position of target human. The same applies to Kalman filter in LT, only instead of using the average of all points inside Final ROI to calculate the position of target human, the average of leg segments (Laser ROI) points inside Final ROI is used. Once the first target human position is confirmed with Vision Human detection result, tracking module *Laser Tracker* is activated. A search window is initialized of size 1m × 1m and with center being the located target human position (Figure 9). Due to the fact that the patterns of human legs can be various and irregular, the most reliable feature is that they are formed of either 1 (legs overlapping) or 2 (legs apart) laser segments with limited segment lengths. Therefore, search window is looking for a complete set of 1 or 2 laser segments in the neighbor area of previously detected human location using prediction of Kalman filter. Once Laser Tracker is activated, LT conducts continuous tracking with search window and the fast tracking performance is supervised by robust FT tracking when the Vision Human detection result is available.

Figure 9 illustrates how search window reacts to the dynamic environments. For the illustration of the size decrease, a scenario is considered where the target human is near a table at t_i. At t_{i+1}, 3 laser segments are detected instead of 2 as besides leg segments, table leg is also included. In this case at t_{i+1}, search window decrease its size until only 1 or 2 laser segments are detected again to avoid potential false tracking. If from $t_{i+1}$to t_{i+2}, target human suddenly increases walking speed so that both legs are out of the search window and no laser segment is detected, the window size increases until at least 1 laser segment is included to avoid losing track of target human.

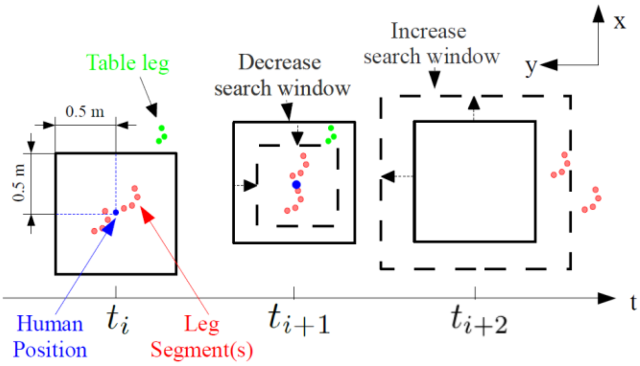


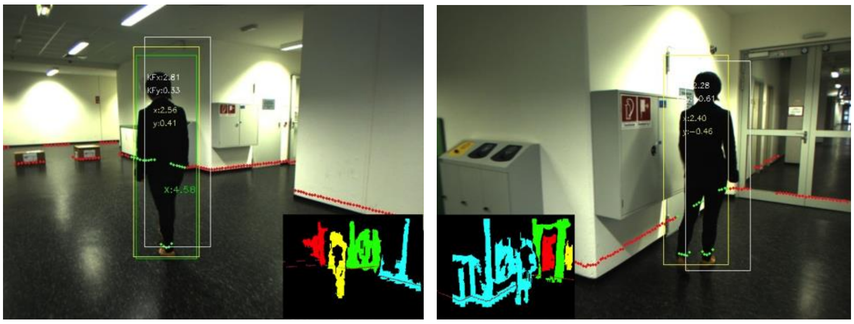
Fig. 9 Decreasing and increasing size of search window

6. IMPLEMENTATION AND EVALUATION

The robot platform used is CORBYS Demonstrator II [17]. It consists of a mobile platform Neobotix MP-M470, a stereo camera Bumblebee XB3 from Point Grey and a 2D LRF by SICK.

6.1. Human detection

As explained in section 3, the human detection performance of Vision Human heavily relies on segmentation result of disparity map, which result in low robustness during tracking. By using low-level fusion data in Fusion Human, the robustness of human detection is significantly improved. Figure 10(a) presents the scenario when the segmentation results are good (as illustrated with different color in small image), therefore both Vision Human (green rectangle) and Fusion Human (green dots) are able to detect target human. However, when target human is close to background and merged into the same blob as background during segmentation (Figure 10(b)), Vision Human fails to detect target human but Fusion System is able to detect target human which is denoted with green dots. In both images in Figure 10, yellow rectangular is visualization of correct Fusion Human result. Namely, starting from green dots an rectangular of fixed size is created. White rectangular in both images is the result of FT Kalman based human prediction.



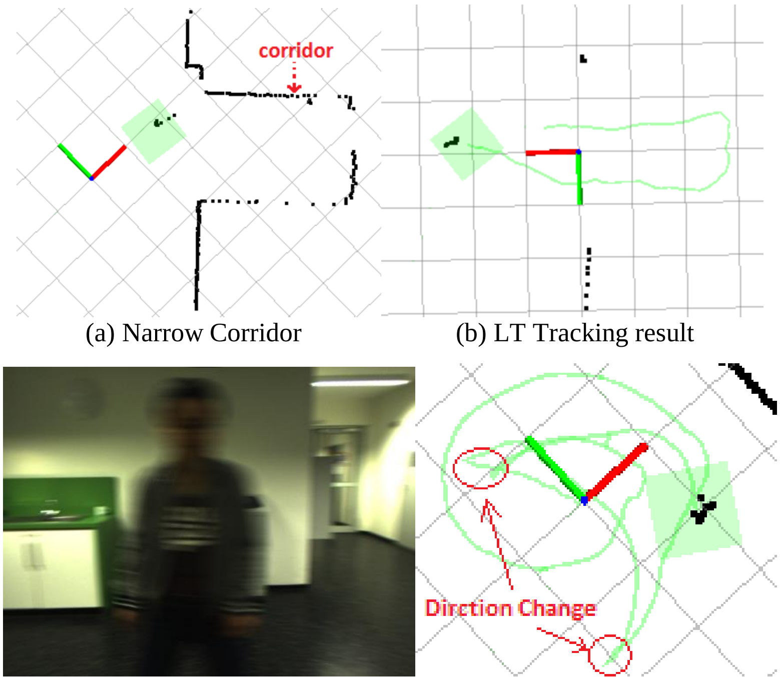
(a) Good Vision Segmentation (b) Bad Vision Segmentation

Fig. 10 *Fusion Human* detection with different vision segmentation results

6.2. Human tracking

In addition to Fusion Human benefits described above (section 6.1), with Fusion Human and Laser Tracker, the proposed Fusion System is able to overcome problems of Vision Human not being able to conduct human detection when target human is partially viewed by vision as well as not being able to conduct fast human tracking since the FOV of stereo camera is narrow.

Figure 11(a) presents the corridors perceived by LRF , the axes denote the robot center, and the green squares are the search windows which include target human. Figure 11(b) shows the LT tracking trajectory of target human after entering and leaving the corridor. Another scenario, which can present the agility and timely tracking performance, is when target human is walking in circles around robot with sudden direction changes. Figure 11(c) shows the stereo camera's view of sudden direction changes and Vision Human loses track of target human. Figure 11(d) presents LT tracking trajectory for this scenario in which the direction changing moments are denoted with red circles. In this case, robot is still capable of tracking target human by using LT.



(a) Narrow Corridor (b) LT Tracking result

(c) Rapid Direction Change (d) High Responsiveness of LT

Fig. 11 *Fusion System* tracking performance

6.3. Performance evaluation

The comparison of performances of Vision Human module and Fusion Human module proposed in this paper was conducted for three indoor scenarios. The first scenario is robot to follow target human in a straight line and second scenario is to follow the target human in and out of a corridor (Figure 11(a)). In the third scenario, target human keeps a close distance to background wall during the whole walking process.

As shown in Figure 12, for the first scenario both modules have the same detection rate, 98.5% (human detected in 134 frames out of total 136 frames). In comparison, the performance of Fusion Human in scenarios two and three shows more advantage. In scenario two, due to the small distance between human and robot in the corridor, Vision Human failed to detect target human for approximately 60 consecutive frames and yields a detection rate of 78.6% (206 frames detected out of total 262 frames). Fusion Human presents a detection rate of 99.2% (260 frames detected out of total 262 frames). In scenario three, the performance of Vision Human degraded more since target human is always close to background wall, which results in a detection rate of 43.3% (153 frames detected out of total 353 frames). In contrast, the performance of Fusion Human shows high-robustness with detection rate of 99.4% (351 frames detected out of total 353 frames).

Figure 13 illustrates real-time capability of proposed method. Due to limited computational capacities of on- board PC, Vision Human (FT) conducts human detection and tracking at 4Hz.By using laser data-based searching window and Kalman prediction, Laser Tracker enables the robot to perform human tracking at 20 Hz. In this way, the responsiveness of tracking process is dramatically increased.

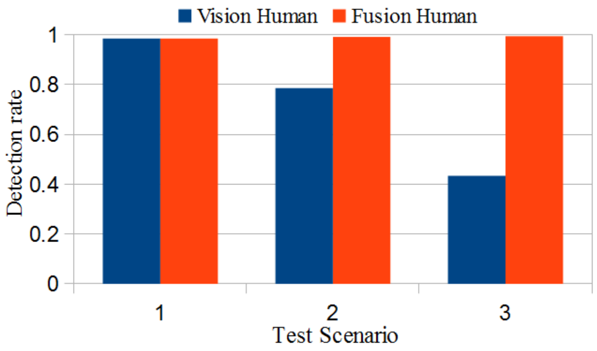


Fig. 12 Performance of *Fusion Human* and *Vision Human* on human detection with 3 scenarios

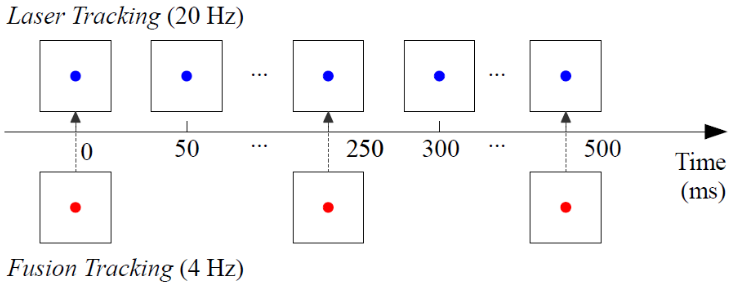


Fig. 13 Performance of LT and FT on human tracking

Acknowledgement: *This research was supported by the European Commission as part of the CORBYS (Cognitive Control Framework for Robotic Systems) project under contract FP7 ICT270219.*

REFERENCES

[1] S. Chen, "Kalman filter for robot vision: A survey," IEEE Transactions on Industrial Electronics, vol. 59, no. 11, pp. 4409 – 4420, 2012. [Online]. Available: http://dx.doi.org/10.1109/TIE.2011.2162714

[2] J. Lee, W. Yu, J. Hwang, C. Kim, "A lazy decision approach based on ternary thresholding for robust target object detection," in *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 3924–3929, 2014. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2014.6907428

[3] S. Kai, H. Yuan, C. Li, "Sensor fusion based human detection and tracking system for human-robot interaction," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4835–4840, 2012. [Online]. Available: http://dx.doi.org/10.1109/IROS.2012.6386222

[4] G. Gate, A. Breheret, F. Nashashibi, "Fast pedestrian detection in dense environment with a laser scanner and a camera," in *Proceedings of IEEE 69th Vehicular Technology Conference*, pp. 1–6, 2009. [Online]. Available: http://dx.doi.org/10.1109/VETECS.2009.5073555

[5] E. Petrovic, A. Leu, D. Ristić-Durrant, V. Nikolić, "Stereo vision-based human tracking for robotic follower," International Journal of Advanced Robotic Systems, 2013. [Online]. Available: http://dx.doi.org/10.5772/56124

[6] C. Chou, J. Li, M. Chang, L. Fu "Multi-robot cooperation based human tracking system using laser range finder," in *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 532 – 537, 2011. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2011.5980484

[7] L. Spinello, R. Triebel, R. Siegwart "Multi-modal people detection and tracking in crowded scenes," in *Proceedings of the Twenty/Third AAAI Conference on Artificial Intelligence*, 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.4095

[8] G. Gao, Calibration of stereo camera and 2d laser scanner for environment perception. Technical report, Institute of Automation, University of Bremen, 2014.

[9] R. Siegwart, I. Nourbakhsh, D. Scaramuzza, *Introduction to Autonomous Mobile Robots*. Massachusetts Institute of Technology, 2011.

[10] T. Tao, J. C. Koo, H. R. Choi, "A fast block matching algorithm for stereo correspondence," in *Proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, pp. 38– 41, 2008. [Online]. Available: http://dx.doi.org/10.1109/ICCIS.2008.4670774

[11] S. K. Natarajan, D. Ristic-Durrant, A. Leu, A. Graeser, "Robust stereo-vision based 3D modeling of real-world objects for assistive robotic applications," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pp. 786–792, 2011. [Online]. Available: http://dx.doi.org/10.1109/IROS.2011.6094716

[12] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Transaction on Information Theory, 1962. [Online]. Available: http://dx.doi.org/10.1109/TIT.1962.1057692

[13] R. Hartley, *A. Zisserman, Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004

[14] G. P. Zhang, "Neural networks for classification: a survey," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 30, no. 4, pp. 451–462, 2000. [Online]. Available: http://dx.doi.org/ 10.1109/5326.897072

[15] R. Faragher "Understanding the basis of the kalman filter via a simple and intuitive derivation," IEEE Signal Processing Magazine, vol. 29, no. 5, pp. 128 - 132, 2012. [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.2203621

[16] G. Welch, G. Bishop, An introduction to the Kalman filter. Techn. Rep., Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC. 1995. [Online]. Available: http:// http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.6578&rep=rep1&type=pdf

[17] CORBYS-Cognitive Control Framework for Robotic Systems, [Online]. Available: www.corbys.eu