The 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2011)
Nov. 23-26, 2011 in Songdo ConventiA, Incheon, Korea

Real-Time Human Body Motion Estimation Based on Multi-Layer Laser Scans

Wei Wang[1], Dražen Brščić[2], Zhiwei He[1], Sandra Hirche[1] and Kolja Kühnlenz[1]

[1]Institute of Automatic Control Engineering, Technische Universität München, D-80290 München, Germany
(Tel : +49-89-289-25725; E-mail: wangwei@lsr.ei.tum.de, zhiwei.he@mytum.de, hirche@tum.de and koku@tum.de)
[2]Advanced Telecommunication Research Institute International (ATR), 619-0288 Kyoto, Japan
(Tel : +81-774-95-1415; E-mail: drazen@atr.jp)

Abstract - Real time human body motion estimation plays an important role in the perception for robotics nowadays, especially for the applications of human robot interaction and service robotics. In this paper, we propose a method for real-time 3D human body motion estimation based on 3-layer laser scans. All the useful scanned points, presenting the human body contour information, are subtracted from the learned background of the environment. For human contour feature extraction, in order to avoid the situations of unsuccessful segmentation, we propose a novel iterative template matching algorithm for clustering, where the templates of torso and hip sections are modeled with different radii. Robust distinct human motion features are extracted using maximum likelihood estimation and nearest neighbor clustering method. Subsequently, the positions of human joints in 3D space are retrieved by associating the extracted features with a pre-defined articulated model of human body. Finally we demonstrate our proposed methods through experiments, which show accurate human body motion tracking in real time.

Keywords - Human body motion estimation, multi-layer laser scans, iterative template matching for clustering

1. Introduction

Nowadays 3D human body motion capturing has been given increasing attention, due to its widespread use in human robot interaction, the analysis of human social behavior and other applications for service robots. Several challenges are still standing, such as the fusion of the limited information from the sensors, useful information extraction.

Multiple cameras and Time of Flight (TOF) camera based approaches are typical in 3D human body motion on-line capturing, due to the obtained richness of information [1], [2], [3]. With the development of the sensor technology, some new sensors are increasingly used within robot perception system, such as the Kinect[1], which can provide full 3D view depth and color information.

Nevertheless, an important issue with these kinds of sensors is that they are quite sensitive to illumination and other environmental changes, which is why most of them are still limited to indoor applications. Compared to these sensors, laser range finders (LRF) are much robuster to illumination changes [4], [5], especially in outdoor applications such as autonomous city explorer robot ACE[2], also provide large scan range, a high data rate, and accurate measurement. However, the disadvantage of standard LRF is that the obtained information is 2D, which is limited comparing to the aforementioned sensors.

There are two ways to overcome the 2D limitation of LRF: actuation or use of multi-layer scanning. With actuating LRFs, as demonstrated in [6], [7], it is possible to achieve very good 3D scans of static environments, but making a full scan is usually quite time-consuming, so that cannot be used for real-time tasks such as human tracking. While multi-layer laser scanning does not have this problem, it can be obtained either using multiple single-layer LRFs or a sensor with built-in multi-layer scanning, such as the ibeo-LUX[3].

Multi-layer LRF systems have been used previously for accurate people detection and tracking [4], [5]. Mozos et al. [4] use a static 3-layer LRFs system to detect the surrounding people. The approach is composed of a probabilistic combination of the outputs from different classifiers which are extracted for each layer to detect a particular body part such as head, torso or leg. A mobile robot with a 2-layer LRFs system is used by Carballo et al. [5], where a human model is built for the association of the different features, e.g. chest and legs areas and a volume representation, which allows the estimation of the current person's position. However, these works only focus on people tracking. There are also some works on estimating both human position and pose using multiple LRFs [8], [9]. Glas et al. [8] propose a people tracking method using particle filter to get not only the people location but also the body rotation and arm position, by using the contour information from the torso-level lasers. Matsumoto et al. [9] try to use four corner LRFs at same height to get the contour features for multiple people pose estimation. Some pose candidates are weighted after the re-sampling step and propagation by the transition model to get the winner pose as the result. The estimated region is fixed and just some poses can be estimated. In the above related works multi-LRF setups are only used for either people location tracking or general pose estimation. As LRF gives a fewer amount of information when compared to other full 3D view sensor, it is challenging to estimate the full human body motion in real time.

The contribution of this paper is an approach to estimate the 3D human body motion in real-time using multi-layer laser range finders which does not only provide the

[1] http://www.primesense.com

[2] http://www.ace-robot.de/

[3] http://www.ibeo-as.com/

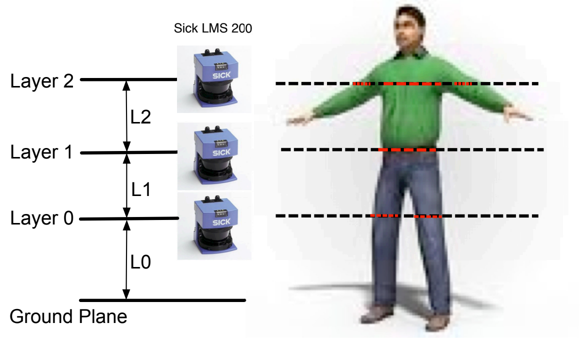


Fig. 1 Overview of the proposed estimation system.

accurate human position but also the full human body pose. The system is presented in Figure 1, where three LRFs are vertically aligned on different fixed heights from the ground plane. The heights are chosen in such a way so that the LRFs can capture the arc-shaped contour points of the human's torso and upper arms, the hip and forearms, and the thighs separately. During segmentation and clustering for these useful scanned foreground points which are extracted from the learning background, a novel iterative template matching method is proposed to solve the self occlusions to get robust distinct human motion features. The system is able to estimate full human body motion after associating the extracted features with the pre-defined articulated human model in real-time.
The remainder of this paper is organized as follows: Section 2 provides the detailed structure of the whole estimation system, including background subtraction, feature extraction, human modeling and data association. The hardware setup and experimental results are presented and discussed in Section 3.

2. Proposed Approach

In this section, we provide the details about the human body motion estimation system resulting in real-time estimation of body joint position in 3D space. The proposed method aims at the problem to get the real human motion based on the limited spatial data in terms of the scanned contour information based on 3-layered laser scans. The processing modules and components are illustrated in Figure 2. After the background subtraction for the raw data from each LRF is done, all the related human arc-shaped contour points are estimated. A novel segmentation and clustering method to extract the human contour features is proposed. The final human pose will be retrieved by associating these features with a predefined articulated human model.

2.1 Background Subtraction

All the scanned contour points of objects in front of LRFs represent the raw data, which means they include both the background information (such as walls and other static objects) as well as the moving objects information. Background information subtraction is needed in order to extract the human contour information from the data.

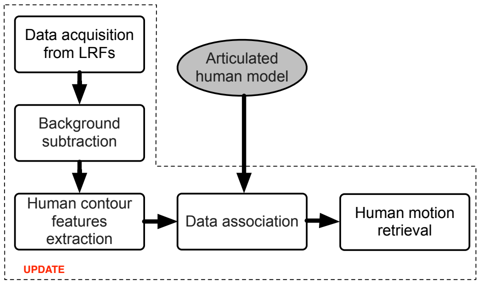


Fig. 2 Diagram of the proposed estimation system.

The three LRFs measurements are with X and Y axes parallel to the ground plane and with heights z_0, z_1 and z_2. All scanned points can be represented as $P = \{p_i\}$ and $p_i = (x_i, y_i, z_i)$, where i is the point index.
The background is learned in the initial stage before the target enters into the scene. We take S scans to average the background information and restore them into the background points set as $P_b = \{\sum_{s=0}^{S} p_s/S\}$. Therefore, the foreground data P_f can be extracted by comparing the raw data with P_b with a given threshold.
However, it is possible that the background changes during the observation, for example due to moved furniture, so the background needs to be dynamically updated. In our strategy, we deal with this case by updating the foreground point set into background data if their position variations are under a given threshold [12].

2.2 Human Contour Features Extraction

From the foreground data P_f, these useful features can be extracted which represented as $F(c, \theta, l)$, where c is the cluster center position, θ is the rotation angle and l is the length of cluster. The information c and θ will be used to get the pose and l will be used to classify these human body parts. Nearest neighbor and template matching these two kinds of segmentation and clustering methods are used to obtain these needed features.

A. Segmentation by Nearest Neighbor Clustering (NNC)

This segmentation criterion is based on the geometry relationship between the nearest neighbor points [10]. All the close enough points will be segmented as one cluster. The clusters with given conditions will be viewed as the effective human contour features.
The algorithm for cluster segmentation is as follows:
1. Compute the distances D between each pair of consecutive points from the effective human data $P_h = \{p_i\}$ in LRF data image: $D_i = \|p_i - p_{i-1}\|$.
2. Classify the suitable points into one cluster $C_j(P, n)$ with vector of points P and the number of points n using the distance threshold T_c: push p_i into C_j if $D_i < T_c$ otherwise create a new cluster.
3. Delete the cluster C_j with index j if its number of points n_{C_j} is under a number threshold T_n as $n_{C_j} < T_n$.
4. Computer C_j center position as feature F_k position information $c_{F_k} = (\sum x_{C_j}/n_{C_j}, \sum y_{C_j}/n_{C_j})$.
5. Compute the rotation and the length of F_k based on the start and end point (p_S, p_E) of the cluster C_j: $\theta_{F_k} =$

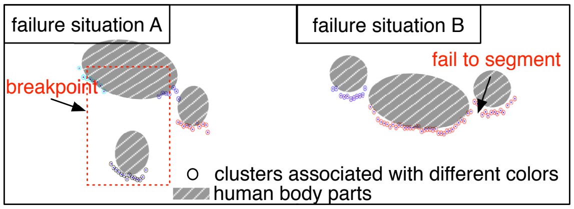


Fig. 3 Two failure situations using NNC. Different colored points show distinct associated clusters.

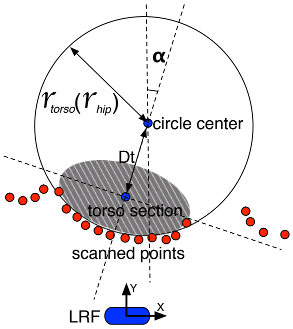


Fig. 4 Torso circle template construction. The red points are scanned points. The real position of torso is represented by the shadow ellipse. The center of the torso can be obtained from the matched circle center position, rotation angle α and the constant distance parameter D_t.

$\arctan \frac{y_{p_S} - y_{p_E}}{x_{p_S} - x_{p_E}}$ and $l_{F_k} = \|p_S - p_E\|$.
However, this geometric clustering method fails when extracting the valid features in two typical situations as shown in Figure 3:

A. Occlusion appears and occluded template cannot be estimated from available estimation, e.g. forearm separates hip part cluster;

B. The outliers of the template can not be excluded which will influence the template estimation result, e.g. when upper arm is within distance threshold T_c to torso.

The final result generated from those estimated clusters in such situations will contain incorrect features. Consequently, template matching is considered to solve the problem of segmentation and associate the related information for clustering.

B. Segmentation Using Template Matching

Aiming to avoid the above failure situations, a circle template matching algorithm based on [13] is employed with the following assumptions: 1) the torso and hip contour are always scanned; 2) the shapes of torso and hip sections in terms of the contour information are not changing.

The torso and hip circle template models are built for matching, where each template has a different radius r_{torso} and r_{hip} , and the circle center has a constant distance D_t to the center of torso section, as shown in Figure 4. The rotation angle α can be obtained from the 5th step of the NNC algorithm in Section 2.2.A. These two circle models radii are defined as $r_{torso} = 320mm$ and $r_{hip} = 300mm$ respectively.

If scanned points can be matched with the circle template, the points should fulfill the following equation

$$Dist = \sqrt{(x - x^*)^2 + (y - y^*)^2} - r = 0. \tag{1}$$

where (x^*, y^*) is the center of the circle template, r is the radius and $Dist$ is the distance between the point and its nearest circle border.

In order to obtain the center position of torso, the circle center needs to be computed first. Therefore, maximum likelihood estimation (MLE) is applied here to estimate the center position of circle. Each scanned 2D laser point is assumed as having an independent error which can be represented as a Gaussian distribution with zero mean and standard deviation σ.

As $(\overline{x}_i,\ \overline{y}_i)$ is the true position of the scanned position (x_i,y_i) and n is the number of scanned points used to matching the circle template, then the likelihood of all the points can be represented as

$$L(x, y) = \prod_{i=1}^{n} \left(\frac{e^{-\frac{(x_i - \overline{x}_i)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{e^{-\frac{(y_i - \overline{y}_i)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \right). \tag{2}$$

For easier computation, we use $-\log L$ to minimize as

$$L_{-log} = -\log\left(\frac{1}{(2\pi\sigma^2)^n} e^{-\frac{\sum_{i=1}^{n} [(x_i - \overline{x}_i)^2 + (y_i - \overline{y}_i)^2]}{2\sigma^2}} \right). \tag{3}$$

By removing all the constants, which do not contribute to minimization, we obtain an equivalent formulation optimization problem with the modified cost function

$$L_{circle} = \sum_{i=1}^{n} \frac{(x_i - \overline{x}_i)^2 + (y_i - \overline{y}_i)^2}{\sigma^2}. \tag{4}$$

The Lagrange method of undetermined multipliers are used including the equality constraint given by Equation (1). The final equation to be used to get the MLE result is

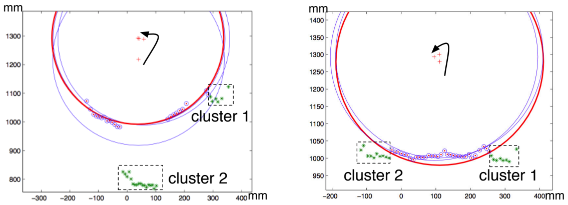
$$L_{circle} = \sum_{i=1}^{n} \frac{[(x_i - x^*)^2 + (y_i - y^*)^2 - r^2]^2}{(x_i - x^*)^2 + (y_i - y^*)^2}. \tag{5}$$

Since the r is the parameter of circle template, MLE becomes the nonlinear problem of obtaining the parameters (x^*, y^*) for minimizing Equation (5). The Newton-Raphson (NR) method is adopted here to solve the optimization problem numerically.

In addition, in order to obtain stable and accurate features, we propose an Iterative Template Matching for Clustering (ITMC) method here. This method can estimate the known template and other clusters whenever the occlusion happens. The pseudocode for ITMC is listed in Algorithm 1.

Algorithm 1 ITMC for segmentation and clustering

```
$i = 0$; //iteration times for ITMC
$T_{outlier}$; //distance threshold to segment the outliers
$\varepsilon_{itmc}$; //threshold for ITMC
$R_{circle}$ //radius of circle template
$//initialize\ circle\ center\ position$
$CE_0 = (x_0^*, y_0^*) =$ center of all points $P = \{p_i\}$;
do
        $i++$;
        $//circle\ matching\ using\ MLE$
        $CE_i^* = (x_i^*, y_i^*) =$ MLE$(P, CE_{i-1}, R_{circle})$;
        $//each\ point's\ distance\ to\ the\ circle$
        $for\ (j = 1; j <= P.size; j++)\{$
                $D_j =$ abs$(\|P_j - CE_i\| - R_{circle})$;
                $if\ D_j > T_{outlier}\ push\ P_j\ into\ O;\}$
        $//reduce\ the\ outliers\ to\ update\ the\ input\ data$
        $P_m =$ all points of $O$;
        $P = P - P_m$;
        $//displacement\ of\ estimated\ circle\ position$
        $\varepsilon = \|CE_i - CE_{i-1}\|$;
while $(\varepsilon \geq \varepsilon_{itmc})$
$//segment\ and\ cluster\ based\ on\ NNC$
$C = \{C_i\} =$ NNC$(O, CE_i) + cluster\{P\}$;
```



(a) (b)

Fig. 5 Clustering based on ITMC. The final matched circle is red. Black arrow shows the trace of matched circle's center during the iterative steps. (a) use 4 times iteration to get the human hip circle template's center, meanwhile segment the forearm clusters; (b) use 3 times iteration to get the human torso circle template's center, meanwhile segment the upper arm clusters.

The main idea of ITMC is to update the input data every time to match the circle and exclude the points which are not related to the circle template. When the position of the matched circle becomes stable, NNC is applied to segment and cluster the excluded points. This algorithm can greatly improve the accuracy of segmentation and clustering results also solve the problem of failure situations shown in Figure 5. Notice that ITMC is employed on layer 1 and layer 2 which used the torso and hip section template, whereas only NNC is applied on layer 0, since it can achieve these clusters for legs feature extraction successfully. Subsequently, all human body features are extracted in real time based on three LRF's data.

2.3 Human Modeling and Data Association

The above extracted features $F(c, \theta, l)$ represent the 2D data at different heights. As the three layered lasers have the fixed height above the ground plane, the system is able to estimate the particular person currently. To get 3D human joints data, the features need to be associated with the pre-defined articulated human model (illustrated in Figure 6), which has 11 fixed length links with 25 degrees of freedoms. Details of the pre-defined human model are shown in Table 1. These coefficients and the radii of torso and hip templates mentioned in Section 2.2.B are measured from the people who is proposed motion estimation target.

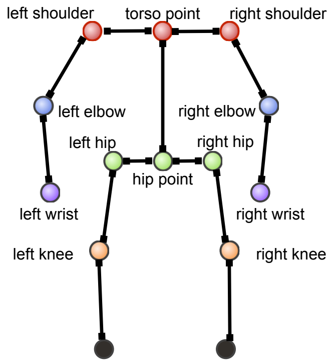


Fig. 6 Articulated human model for data association.

Table 1 Parameters of human model.

Link	Start Joint	End Joint	Length(mm)	DOFs
1	hip point	torso point	490	3
2	torso point	left shoulder	210	1
3	torso point	right shoulder	210	1
4	left shoulder	left elbow	290	3
5	right shoulder	right elbow	290	3
6	left elbow	left wrist	320	3
7	right elbow	right wrist	320	3
8	hip point	left hip	200	1
9	hip point	right hip	200	1
10	left hip	left knee	500	3
11	right hip	right knee	500	3

The human contour features f_{torso} and f_{hip} have been obtained by the proposed ITMC approach, while all the other features can be classified based on the position relationship between each layer height, in the particular, right and left arms are classified depending on clusters center position. Classified features then can be associated with the corresponding parts of human body as $F_{body} = \{f_{torso}, f_{hip}, f_{l-upperarm}, f_{r-upperarm}, f_{l-forearm}, f_{r-forearm}, f_{l-thigh}, f_{r-thigh}\}$. Each feature just represents the horizontal sectional center position of associated human body part. Note that the position data of F_{body} is a 2D position with fixed height. Consequently, the next step is to get each related human joint position in 3D space based on these extracted features and retrieve the human body motion in real time.

In order to simplify the association of features, the height of hip in terms of $f_{l-thigh}$ and $f_{r-thigh}$ is fixed, which is a reasonable approximation as long as the human is standing or walking only[11]. Furthermore the links 2,3,8 and 9 are approximated to have only one DOF, as we ignore the vertical rotation of the torso and hip. In addition, the foot joints are also neglected because of the lack of related height information. With these assumptions, all associated human body part features F_{body} are extended into related human joint data with hierarchical computation as shown in Figure 7:

1. Based on the torso and hip features' information, get the hip point with the fixed height;

2. Find the torso joint based on the fixed length of torso

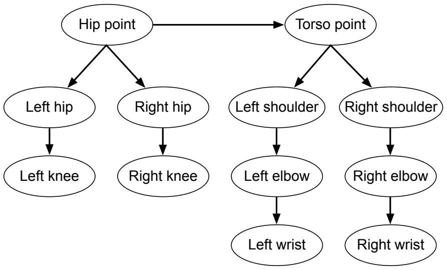


Fig. 7 Hierarchical computation for each joint

and direction with torso and hip features;

3. Compute the rest of the joints hierarchically.

Finally, all predefined human joints positions in 3D space can be estimated as P_{joints}. The position data are filtered by Kalman filter in order to get smoother result.

3. Experimental Results

The experimental setup consists of 3 SICK LMS-200 laser range scanners from SICK AG[4]. The set of LRFs at each layer can scan with distance resolution of 10mm and angular resolution of 0.5 degree, angular range of 180 degree, and data transmission rate at 500kBps using the RS-422 interface. In the experiments, the three LRFs have been fixed to the heights from 590 mm, 950 mm and 1255mm as the vertical distance to the ground plane.

The synchronization of data from the LRFs is not a problem due to the relatively fast scanning rate and careful mounting also alleviates the need for additional calibration. We use the Sick LIDAR Toolbox and the processing in all estimation steps is done in real time using Matlab on a Core Duo PC (Linux kernel x86). The processing time is below 40 ms, which is fast enough to estimate the human motion. While standing in front of LRFs one human perform several motions, such as walking, running, arm waving and moving sideways.

In order to better evaluate proposed system, a Kinect sensor is used as a reference, by running the human pose tracking software from the Kinect middleware. The experimental results of this 3-layer LRF estimation system and tracking system by Kinect are shown Figure 10. The results show that, our proposed system achieves accurate full body motion, while in the effective range of Kinect, the system performs very close to the results obtained with Kinect. But one thing should be noted, Kinect sensors angular range is only 60 degree, while 3-layer LRF systems is 180 degree. As shown in Figure 8, our estimation system still can perform quite well in the location with larger angle, while the Kinect sensor does not provide any results anymore. Furthermore, the speed of our proposed method is 25Hz while Kinect's is 13Hz. All in all, despite of the fewer data information of LRF, a relatively good estimation result could be obtained within the area of half circle (180 degrees) with 4m radius.

For the data accuracy evaluation, a Phoenix Technologies Incorporated VZ-4000 3D position motion sensing system[5](PTI) is used to capture the joint's position as ground truth. Because of the limitation of the tracking area of this system, only the right arm's motion are analyzed here, which is described by the right wrist's 3D position $p_{r-wrist} = (x, y, z)$, the angle β between the torso and upper arm and the angle γ between upper arm and forearm. These data can fully describe the pose of the right arm. The estimation is evaluated during dynamic motions. The performance of position and angle are shown in Figure 9. Table 2 gives the evaluation of root mean square deviation (RMSD) during the experiment. As is obvious from the results, the estimated arm pose is very close to the measurement data from PTI. The errors in z dimension and in the estimated arm angles are somewhat larger, which is mainly due to the redundancy of the model and the approximate model parameters. The jumps during the estimation are the result of occasional mismatching of features in the scan.

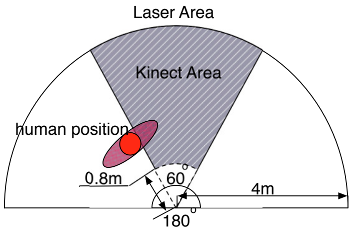


Fig. 8 Different effective area of Kinect sensor and laser scanners.

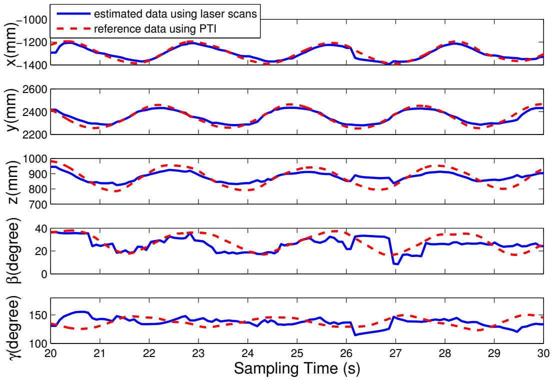


Fig. 9 Measured and estimated pose of right arm during dynamic motion-back-forth swinging of the arm. Pose is described with wrist position (x, y, z) , the angle β and the angle γ. Estimated data using laser scans is shown with full blue line and reference data using PTI is shown with red dashed line.

4. Conclusion and Future Work

In this paper, we propose a real time human body motion estimation system based on 3-layer laser range finder scans. To solve the problems of segmentation and clustering in some problematic poses, we used iterative template matching for clustering method (ITMC) to get robust ex-

[4] http://www.sick.com

[5] http://www.ptiphoenix.com/

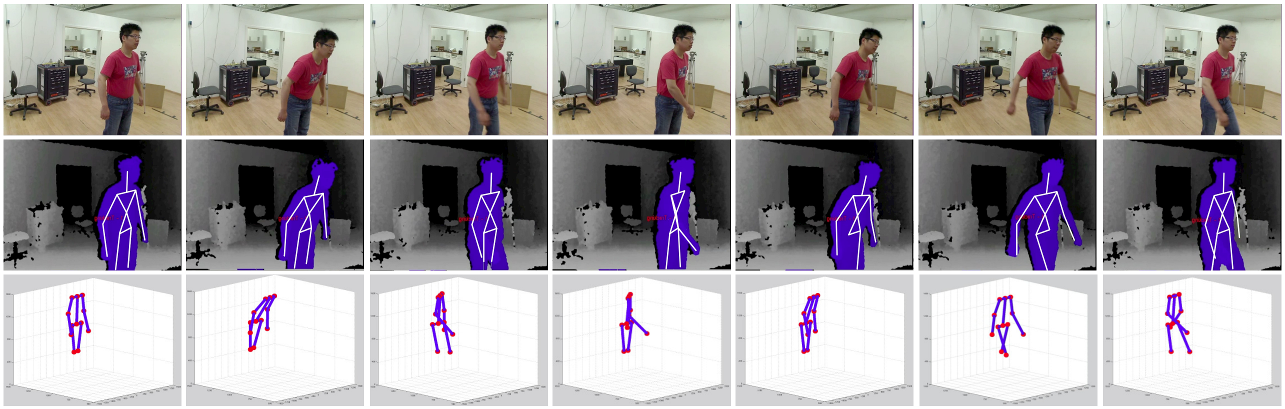


Fig. 10 Human body motion estimation using OpenNI user tracker from Kinect (middle figure) and using the proposed 3-layer LRFs (bottom row). The top row shows the image reference.

Table 2 RMSD for the right arm pose

position error (mm)			angle error$(degree)$	
x	y	z	β	γ
21.63	22.79	33.43	6.12	11.84

traction of body parts. In our experiment, the result is based on the particular human model because of the limitation of the source data and the fixed height of sensors. For the single specific human pose estimation, our approach cannot only retrieve the human motion but also the accurate 3D position of human joints.
Future work will focus on the multiple person motion estimation with different articulated human models.

5. Acknowledgement

This work is supported in part by the DFG excellence initiative research cluster ”Cognition for Technical Systems – CoTeSys,” see www.cotesys.org, the FP7 EU-STREP ”Interactive Urban Robot (IURO)”, see www.iuro-project.eu, the Institute for Advanced Study (IAS), Munich.

References

[1] Y. Iwashita, R. Kurazume, T. Mori, M. Saito and T. Hasegawa, Model-based Motion Tracking System using Distributed Network Cameras, IEEE International Conference on Robotics and Automation Anchorage Convention District, Alaska, USA, 2010.

[2] S. Pellegrini and L. Iocchi, ”Human Posture Tracking and Classification through Stereo Vision and 3D Model Matching”, *EURASIP Journal on Image and Video Processing*, ID 476151, 2008.

[3] S. Knoop, S. Vacek, K. Steinbach and R. Dillmann, Sensor Fusion for Model Based 3D Tracking, IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems, Heidelberg, Germany, September 2006.

[4] S. Pellegrini, L. Iocchi. O. M. Mozos, R. Kurazume and T. Hasegawa, ”Multi-Part People Detection Using 2D Range Data”, *International Journal of Social Robotics*, Volume 2, Number 1, 31-40, 2010.

[5] A. Carballo, A. Ohya and S. Yuta, Multiple People Detection from a Mobile Robot using Double Layered Laser Range Finders, ICRA Workshop on People Detection and Tracking, Kobe, Japan, 2009.

[6] M. Matsumoto and S. Yuta, 3D Laser Range Sensor Module with Roundly Swinging Mechanism for Fast and Wide View Range Image, IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems, pp. 156-161, 2010.

[7] A. Nüchter, K. Lingemann, J. Hertzberg and H. Surmann, Accurate Object Localization in 3d Laser Range Scans, 12th International Conference of Advanced Robotics, pp. 665–672, 2005.

[8] D. F. Glas, T. Miyashita, H. Ishiguro and N. Hagita, Laser Tracking of Human Body Motion using Adaptive Shape Modeling, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 603-608, 2007.

[9] T. Matsumoto, M. Shimosaka, H. Noguchi, T. Sato and T. Mori, Pose Estimation of Multiple People using Contour Features from Multiple Laser Range Finders, IEEE/RSJ international conference on intelligent robots and systems, St. Louis, USA, 2009.

[10] K. C. J. Dietmayer, J. Sparbert and D. Streller, Model Based Object Classification and Object Tracking in Traffic Scenes from Range Images, IEEE Intelligent Vehicles Symposium, Tokyo, Japan, pp.25-30, 2001.

[11] T. Ha and C. Choi, ”An Effective Trajectory Generation Method for Bipedal Walking”, *Robotics and Autonomous Systems archive*, Vol 55, No 10, 2007.

[12] D. Brščić and H. Hashimoto, ”Mobile Robot as Physical Agent of Intelligent Space”, *Journal of Computing and Information Technology*, Vol 17, No 1, 2009.

[13] H. Tamura, T. Sasaki, H. Hashimoto, and F. Inoue, Position Measurement System for Cylindrical Objects using Laser Range Finder, IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 2010.