Human Pose Estimation from Depth Image Using Visibility Estimation and Key Points

Sungjin Huh and Gyeonghwan Kim

Department of Electronic Engineering, Sogang University
CPO BOX 1142, Seoul 100-611, Rep. of Korea
{seungjin,gkim}@sogang.ac.kr

Abstract. In this paper, we propose the upper body pose estimation algorithm using 3-dimensional model and depth image. The conventional ICP algorithm is modified by adding visibility estimation and key points - extreme points and elbow locations. The visibility estimation keeps occluded points from participating in pose estimation to alleviate the affection of self-occlusion problem. Introduction of extreme points and elbow locations, which are extracted using geodesic distance map and particle filter, improves the accuracy of pose estimation result. The optimal parameters of the model are obtained from nonlinear mathematical optimization solver. The experimental results show that the proposed method accurately estimates the various human poses with self-occlusion.

Keywords: human pose estimation, 3D model based, modified ICP, self-occlusion, key points, geodesic distance.

1 Introduction

In Human Computer Interaction (HCI), introduction of gesture recognition provides more intuitive and convenient user interfaces than what traditional input devices do such as keyboards and mouses, etc. Since the recognition task is based on human poses, accurate estimation of the pose is an important prerequisite. Many research efforts have been made can be classified into two categories: marker-based methods and vision-based methods using RGB image sensor. The marker-based pose estimation methods which have been used in mainly film industry need burdensome suits or markers. Although estimation of pose using RGB image sensor is always preferred due to its convenience of data acquisition, loss of depth information during projection onto 2D image plane is a major hurdle for the pose estimation. Recently, pose estimation using depth image has been popularized by the Kinect sensor. Depth sensors are robust to variations of visual appearances such as illumination, texture, etc. The advantage of depth sensors and easy access of depth sensors lead vivid research on human pose estimation.

Shotton et al.[1] proposed a method of pose estimation using random forest[2] to segment an observed depth image into distinct body parts. Hemandez-Vela et al.[3] augmented graph-cut optimization to the method of Shotton et al.[1] to

V.G. Duffy (Ed.): DHM/HCII 2013, Part II, LNCS 8026, pp. 333–342, 2013.

improve the segmentation performance. The learning based approaches require large amount of database to cover the variety of human poses, and also have potential dependency on the database. The method proposed by Grest et al.[4] is based on the ICP(iterative closest point) algorithm to fit a body model to the depth image. Plagemann et al.[5] and Schwarz et al.[6] extracted anatomic key points from the depth images. However, depth cues, extensively used by Grest et al.[4] and the sparse key points[5][6] are prone to getting stuck into local minima in depth image and self-occlusion that are common in human pose.

In this paper, we define a 3-dimensional human upper body model and propose a pose estimation scheme by finding the optimal parameters, in order to alleviate the disadvantages of learning-based methods. To keep model parameters from deterioration due to the self-occlusion, visibility of a body part is continuously evaluated. Also, the local minima problem is effectively handled by inclusion of key points – extreme points and elbow joints - into the objective function. The optimal parameters are obtained by solving the mathematical optimization problem, similarly to the conventional ICP[7].

2 Human Pose Estimation Method

Given a sequence of depth images, a scheme of model based human pose estimation is described in this section. The articulated 3D body model is defined and nonlinear optimization is applied to the problem using a modified ICP algorithm. Unlike the conventional ICP algorithm, the modified version contains the visibility term and additional key points. Detailed description on the optimization method of the modified ICP algorithm and estimation of the key points are presented in the following sections.

2.1 Human Upper Body Model

In this paper, human upper body is prototypically modeled as concatenation of head, torso, and left/right arms as in Figure 1. To represent curvature of human body parts, the models of head and torso take on an elliptic cylinder and a cone, respectively. The arm model takes on a folding line to represent the flection of the elbow. The entire upper body model has 17 DOF: the translations of head, torso, and the angles of torso, shoulders and the elbow flection $(\mathbf{t}_{head}, \mathbf{t}_{torso}, \boldsymbol{\theta}_{torso}, \boldsymbol{\theta}_{arm}, \theta_{elbow})$. The sample points of each model of body parts are arranged in a regular grid and exploited to estimate the pose.

The pose of the articulated 3D model is assumed to be described by similarity transformation $S(\boldsymbol{\theta}, \mathbf{t})$, which contains rotation and translation. Figure 2 shows the conceptual projection process of the torso from 3D space onto 2D depth image plane $d(x, y)$, performed by a depth sensor. Since intrinsic and extrinsic parameters of the depth sensor are fixed, the pose is solely determined by $\boldsymbol{\theta}$ and $\mathbf{t}$.

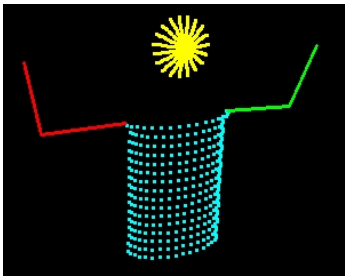


Fig. 1. The proposed 3-dimensional human body model

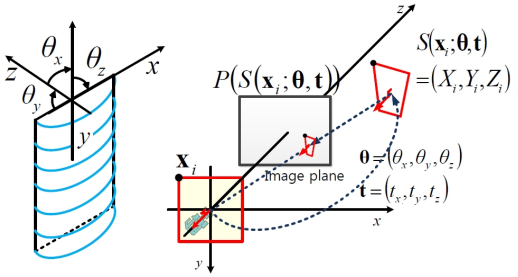


Fig. 2. Conceptual projection process of torso model with similarity transform and projection matrix

2.2 Modified ICP Algorithm

In our implementation, we made modification of the ICP algorithm which has shown successful results in human pose estimation[4][8][9][10][11]. The modified ICP algorithm is to find $\boldsymbol{\theta}^*$ and $\mathbf{t}^*$ which minimize the squared distance between the transformed model points $S(\mathbf{x}_1;\boldsymbol{\theta},\mathbf{t})$ and the corresponding depth measurement $d_i = d(P(S(\mathbf{x}_i;\boldsymbol{\theta},\mathbf{t})))$. The resulting objective function to be minimized has a form of

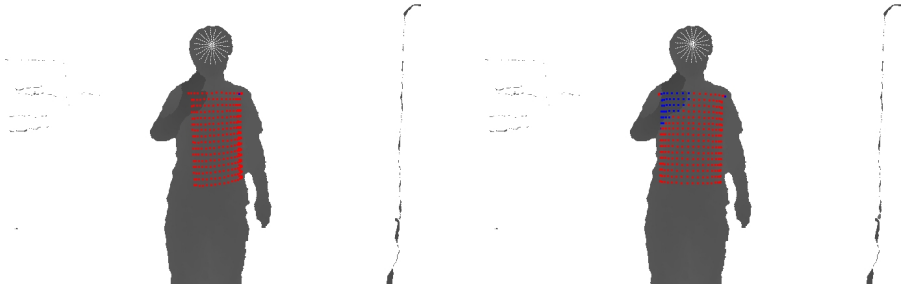
$$J_{ICP} = \sum_{i=1}^{N} g_i \cdot |d_i - Z_i|^2 \tag{1}$$

where $N, \mathbf{x}_i, Z_i$ and g_i represent the number of model points, i-th model point, depth of the transformed model point of $\mathbf{x}_i$ and the corresponding visibility function, respectively.

Self-occlusion occurred in human body pose estimation can be detected by comparing Z_i with the corresponding depth measurement d_i. The visibility function g_i is determined as (2),

$$g_i = \begin{cases} 0, & | d_i - Z_i | < c \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

where c is a criterion for division. If the model point is occluded by another body part, the similarity transformed model point $S(\mathbf{x}_i;\boldsymbol{\theta},\mathbf{t})$ is farther than the threshold c compared to the corresponding depth value d_i, and the g_i is set to be zero. In case self-occlusion occurs, the occluded model point needs to ge ignored in estimating $\boldsymbol{\theta}$ and $\mathbf{t}$. When self-occlusion is not considered, estimation of the parameters is seriously affected by wrong measurement, d_i. Figure 3 shows the role of visibility estimation. In Fig. 3(a), the torso model is heavily affected by self-occlusion and this makes the arm pose not to be estimated properly. Figure 3(b) shows that the problem of self-occlusion can be handled by the visibility estimation. In Fig. 3(b), the sample points on torso model that are occluded by the arm are eliminated from calculating the optimal parameters. Consequently, the pose is estimated properly without occluded points.



(a) Torso pose estimation without visibility estimation:the torso model is affected by self-occlusion and is not estimated properly

(b) Torso pose estimation with visibility estimation:the torso model is estimated properly with visibility estimation

Fig. 3. Role of the visibility estimation

2.3 Key Points from Geodesic Distance Map and Using Particle Filter

The parameter estimation by the conventional ICP tends to be trapped in a local minimum, as shown in Fig. 3(a). For more stable and reliable pose estimation, addition of extra key points[5][6] has been taken into account. In [5][6], they extracted key points based on the geodesic distance from the center of mass of the body. However, if the path from the center of the mass to an extreme point is cut off because of self-occlusion, the position of the extreme point cannot be determined properly. Based on the fact that torso parameters are relatively reliable than the others, locations of shoulders are easily determined. Unlike the method of [5][6], We are able to measure the geodesic distance from shoulders to extreme points even if the path from the center of mass to extreme points are occluded.

Euclidean distance between adjacent pixels $\boldsymbol{u}$ and $\boldsymbol{v}$ in 3D space is given by

$$d_e(\boldsymbol{u}, \boldsymbol{v}) = \left\| \frac{\boldsymbol{u}}{d(\boldsymbol{u})} - \frac{\boldsymbol{v}}{d(\boldsymbol{v})} \right\|, \tag{3}$$

and the geodesic distance between two point $\boldsymbol{s}, \boldsymbol{t}$ is then

$$d_g(\boldsymbol{s}, \boldsymbol{t}) = \sum_{(\boldsymbol{u}, \boldsymbol{v}) \subset P_{\boldsymbol{s},\boldsymbol{t}}} d_e(\boldsymbol{u}, \boldsymbol{v}) \tag{4}$$

where $P_{\boldsymbol{s},\boldsymbol{t}}$ is the shortest path between $\boldsymbol{s}$ and $\boldsymbol{t}$. Considering the initial direction of the path, the extreme points are distinguished from each other. Reducing the distance between the estimated position of extreme point and the tip of the arm model participates in objective function as (5). Figure 4 shows the position of estimated extreme points. The geodesic distance from shoulder is displayed in left column of Fig. 4. By considering the direction of geodesic distance to the

extreme point and using non-maximal suppression, left/right extreme points are extracted as shown in right column of Fig. 4.

Besides the extreme points, the location of elbows takes significant role in estimating arm model parameters. Generally, the arm model is easier to lose the actual arm pose and stuck in local minimum than the torso and the head models because of its relatively large degree of movement of arm.

Therefore, estimating elbow location takes large advantage of improving accuracy. The particle filter is well known as a means for tracking nonlinearly fast moving objects due to its stochastic nature[12]. Kim and Kim[10] proposed a method of tracking limbs using the ICP and the particle filter. In this paper, the elbow location is estimated by particle filter based on distance from shoulder in geodesic distance map.

Particle filter estimates the unknown state $\boldsymbol{e}_t$, which is the location of a particle on the image plane, from the observation $l_{1:t} = \{l_1, \ldots, l_t\}$ of which each element refers the distance from shoulders to the particle. Then the posterior density $p(\boldsymbol{e}_t|l_{1:t})$ are approximated with a sum of N_p weighted particles $\left\{\boldsymbol{e}_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N_p}$ with $\sum_{i=1}^{N_p} w_t^{(i)} = 1$. The general procedure of the particle filter is followed: resampling, prediction, and update steps. Since the elbow location can move arbitrary, the particles are drawn and regenerated by uniform state transition model in prediction step. In update step, the weight of each particle is updated based on the observation likelihood as $w_t^{(i)} \propto p(l_t|\boldsymbol{e}_t^{(i)})$. Figure 5 shows that the elbow location estimation encourages the model to escape local minimum.

The objective function to be minimized is defined as combination of conditions mentioned in the previous sections.

$$\begin{aligned} J &= J_{ICP} + J_{extreme} + J_{elbow} \\ &= \sum_{i=1}^{N} g_i \cdot |d_i - Z_i|^2 + \sum_{j=1}^{2} \left\{\boldsymbol{p}_{extreme,j} - S(\boldsymbol{x}_{extreme,j}; \boldsymbol{\theta}, \mathbf{t})\right\}^2 \\ &\quad + \sum_{j=1}^{2} \left\{\boldsymbol{p}_{elbow,j} - S(\mathbf{x}_{elbow,j}; \boldsymbol{\theta}, \mathbf{t})\right\}^2 \end{aligned} \tag{5}$$

where $\boldsymbol{p}_{extreme}$ and $\boldsymbol{p}_{elbow}$ are estimated position of extreme point and elbow location in 3D space, respectively. Letting $\boldsymbol{\Theta} = [\boldsymbol{\theta}^T, \mathbf{t}^T]^T$, the optimal solution $\boldsymbol{\Theta}^*$ can be obtained by finding the parameters minimizes J as (6).

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} J \tag{6}$$

2.4 Nonlinear Optimization

Since the objective function given in (5) is highly nonlinear due to the nature of human pose and the visibility function g_i, optimization of the objective function is performed in an iterative manner. Though the simplest and straightforward method of optimization is gradient descent[13], i.e. $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \eta \nabla J(\boldsymbol{\Theta})$, excessive number of evaluating the gradient $\nabla J(\boldsymbol{\Theta})$ results in slow convergence. Moreover, difficulties in specifying the value of η is another reason to step aside from the gradient descent method.

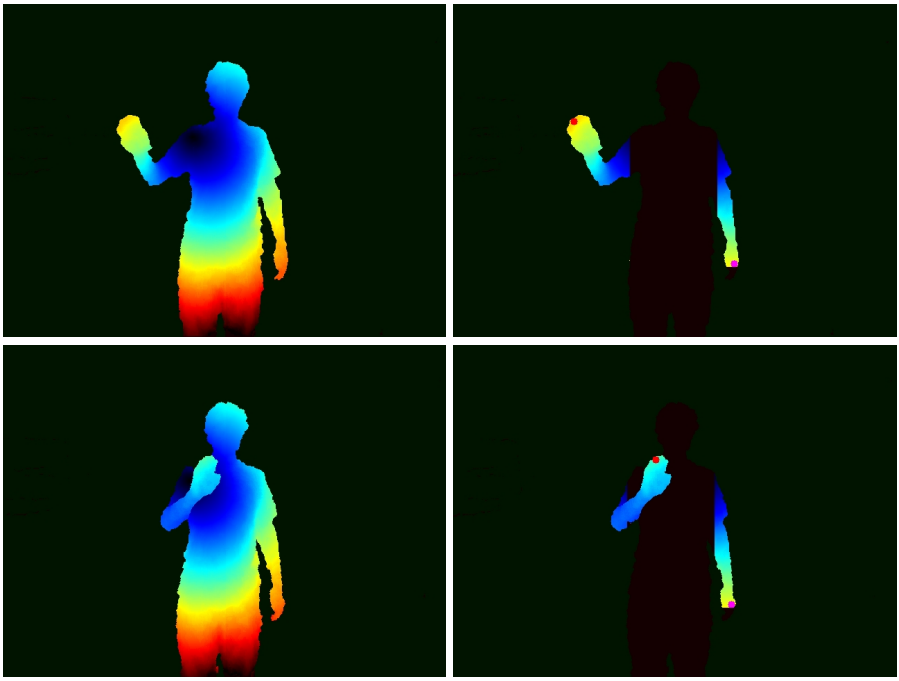
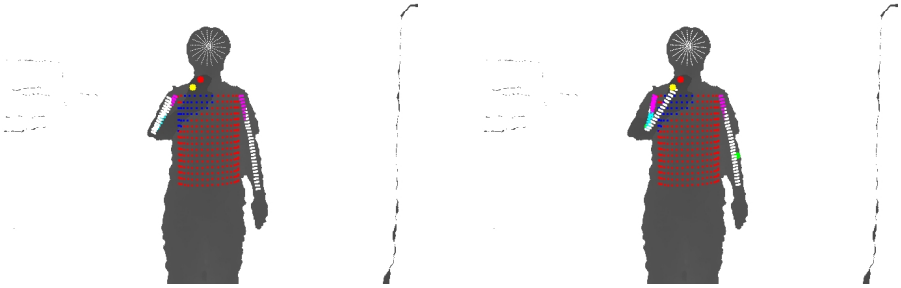


Fig. 4. Estimated extreme points based on the geodesic distance from shoulders: by considering the initial direction, the right hand is distinguished from the left hand



(a) Arm pose without elbow location estimation:the arm model stuck in a local minimum

(b) Arm pose with elbow location estimation:with assistance of elbow location, the arm model escaped a local minimum

Fig. 5. Role of elbow location estimation

Therefore, we employed Gauss-Newton method[13] for the optimization. Gauss-Newton method is one of variations of Newton's method and is widely used for data-fitting problems as (5). For brevity of description, (5) is rewritten in a least square form,

$$J(\mathbf{\Theta}) = \mathbf{r}^T\mathbf{r}, \quad \mathbf{r} = [r_1, \ldots, r_M](M = N + 4) \tag{7}$$

where the residual $\mathbf{r}$ is a column vector containing the scalar components of J_{ICP}, $J_{extreme}$ and J_{elbow}. As given the objective function, the gradient $\nabla J(\mathbf{\Theta})$ and the Hessian matrix $\mathbf{H}$ of $J(\mathbf{\Theta})$ can be represented as,

$$\nabla J(\mathbf{\Theta}) = 2(\mathbf{J}(\mathbf{\Theta}))^T\mathbf{r}(\mathbf{\Theta}) \tag{8}$$

$$\mathbf{H} = 2(\mathbf{J}(\mathbf{\Theta})^T\mathbf{J}(\mathbf{\Theta}) + \mathbf{S}(\mathbf{\Theta})) \tag{9}$$

where $\mathbf{J}(\mathbf{\Theta})$ refers the Jacobian matrix and $\mathbf{S}(\mathbf{\Theta})$ refers the matrix whose (k, j)th component is

$$\sum_{i=1}^{M} r_i(\mathbf{\Theta})\frac{\partial^2 r_i}{\partial\theta_k\partial\theta_j}(\mathbf{x}). \tag{10}$$

Unlike the Newton's method, $\mathbf{r}$ is assumed to be linear in the Gauss-Newton method, i.e. $\partial^2 r/\partial\theta_k\theta_j = 0$. Then, the second derivatives in the Hessian matrix of $J(\mathbf{\Theta})$ vanishes and the Hessian matrix is approximated by

$$\tilde{\mathbf{H}} = 2\mathbf{J}(\mathbf{\Theta})^T\mathbf{J}(\mathbf{\Theta}). \tag{11}$$

The approximation of Hessian matrix in (11) effectively eliminates the well-known instability problem of the Newton's method. At each Gauss-Newton iteration, update of $\mathbf{\Theta}$ is calculated by

$$\mathbf{\Theta}^{(k+1)} = \mathbf{\Theta}^{(k)} - \tilde{\mathbf{H}}^{-1}\nabla J(\mathbf{\Theta}). \tag{12}$$

In spite of the Gauss-Newton method, occasional divergences of solution are observed in our experiments. Therefore further relaxation of iterative optimization is used by employing ω-Jacobi method[14]. Solving (12) by the classic Jacobi method, the approximated Hessian matrix $\tilde{\mathbf{H}}$ is decomposed into a diagonal matrix $\mathbf{D}$ and the off-diagonal matrix $\mathbf{O}$,

$$\tilde{\mathbf{H}} = \mathbf{D} + \mathbf{O}. \tag{13}$$

Difference between the Jacobi method and the ω-Jacobi method is the introduction of an additional relaxation factor ω which controls contribution of a new solution to the previous solution. The ω-Jacobi method is described as

$$\mathbf{\Theta}^{(k+1)} = (1-\omega)\mathbf{D}^{-1}\nabla J(\mathbf{\Theta}) - \omega\boldsymbol{D}^{-1}\mathbf{O}\mathbf{\Theta}^{(k)}. \tag{14}$$

Since $\mathbf{D}^{-1}$ is a diagonal matrix, multiplication of $\mathbf{D}^{-1}$ and vectors in (14) are simply computed by component-wise division.

In the ω-Jacobi method, trade-off between the convergence speed and the stability of iterative solution is largely dependent on the relaxation factor ω. When $\omega > 1$, convergence becomes faster than the standard Jacobi method but suffers from instability, and vice versa. In our implementation, we took $\omega < 1$ to prevent the divergence in the iterative optimization.

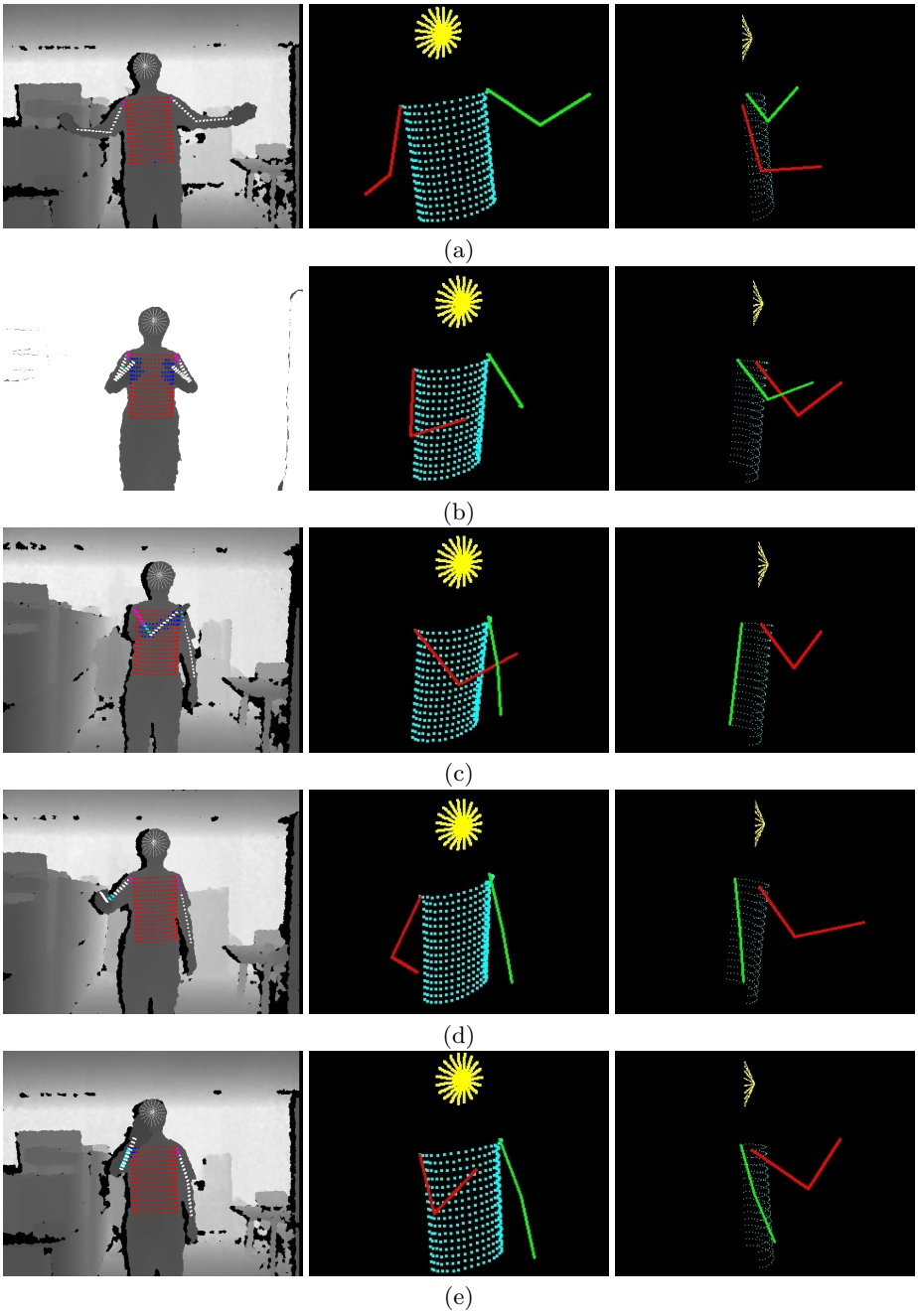


Fig. 6. Human pose estimation results:left column shows the result of pose estimation on depth image. Middle and right columns present the 3D pose corresponding to the left column.

3 Experimental Results

The proposed algorithm is evaluated in various human poses including self-occlusion and the estimation results are described in this section. The microsoft Kinect device was used to take depth images with a resolution of 640x480 pixels. As preprocess procedures, a modified mean filter and a Gaussian filter are applied to depth image for removing noise and smoothing.

Figure 6 shows the pose estimation results on depth images. In the left column, the estimated upper body model is displayed in 2D depth image. Occluded body and arm model points are marked blue and cyan respectively. 3D views of each pose are displayed in center and right columns, respectively. Figure 6(a) presents the accuracy of pose estimation. The model is estimated accurately not only the pose of arm but also the the angle of body. Figure 6(b) and 6(c) shows that the visibility estimation prevents torso model to be affected by occlusion. As stated in the previous section, the torso model points that is occluded by the arm are rejected in estimation of torso model for accuracy. Figure 6(d) and 6(e) illustrate the pose estimation results overcoming the situation that the upper arm is occluded by the lower arm. Not only the visibility function but also well estimated elbow locations and extreme points, the arm pose is estimated properly in spite of occlusion.

4 Conclusion

This paper has proposed a method for estimating human pose from sequences of depth images. We define a 3-dimensional human upper body model and estimate the parameters of the human pose using the modified ICP algorithm with visibility estimation and key points. Our model based algorithm does not require any training data and can estimate arbitrary pose for gesture-based HCI application. Because of visibility estimation, the model based pose estimation overcomes the self occlusion problem by neglecting occluded sample points. Key points obtained from geodesic distance map improve accuracy of arm pose estimation. The experimental results show that our method estimates various upper body poses in accuracy, including self-occlusion.

Acknowledgement. This research was partially supported by the Sogang University Research Grant (SRF-201214005).

References

1. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A.: Efficient human pose estimation from single depth images. IEEE Transactions on Pattern Analysis and Machine Intelligence PP(99), 1 (2012)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

3. Hernandez-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., Escalera, S.: Graph cuts optimization for multi-limb human segmentation in depth maps. In: IEEE Conference onComputer Vision and Pattern Recognition, pp. 726–732 (2012)
4. Grest, D., Woetzel, J., Koch, R.: Nonlinear body pose estimation from depth images. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 285–292. Springer, Heidelberg (2005)
5. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: IEEE International Conference on Robotics and Automation, pp. 3108–3113 (2010)
6. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing 30(3), 217–226 (2012)
7. Besl, P.J., McKay, H.D.: A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(2), 239–256 (1992)
8. Siddiqui, M.: Human pose estimation from a single view point. PhD thesis, University of Southern California, Adviser - Gerard Medioni (2009)
9. Knoop, S., Vacek, S., Dillmann, R.: Modeling joint constraints for an articulated 3D human body model with artificial correspondences in ICP. In: IEEE-RAS International Conference on Humanoid Robots, pp. 74–79 (2005)
10. Kim, D., Kim, D.: A novel fitting algorithm using the ICP and the particle filters for robust 3D human body motion tracking (2008)
11. Droeschel, D., Behnke, S.: 3D body pose estimation using an adaptive person model for articulated ICP. In: Jeschke, S., Liu, H., Schilberg, D. (eds.) ICIRA 2011, Part II. LNCS, vol. 7102, pp. 157–167. Springer, Heidelberg (2011)
12. Haug, A.: A tutorial on bayesian estimation and tracking techniques applicable to nonlinear and non-gaussian processes. MITRE Corporation, McLean (2005)
13. Chong, E.K., Zak, S.H.: An introduction to optimization. Wiley-interscience (2004)
14. Trottenberg, U., Oosterlee, C.W., Schüller, A.: Multigrid. Academic Pr. (2001)