International Conference on Control, Automation and Systems 2010
Oct. 27-30, 2010 in KINTEX, Gyeonggi-do, Korea

Fast 3-D Human Motion Capturing from Stereo Data Using Gaussian Clusters

Nguyen Duc Thang[1] , Tae-Seong Kim[2], Young-Koo Lee[1], and Sungyoung Lee[1]
[1] Department of Computer Engineering, Kyung Hee University, South Korea
(E-mail: ducthang@oslab.khu.ac.kr, yklee@khu.ac.kr, sylee@oslab.khu.ac.kr)
[2] Department of Biomedical Engineering, Kyung Hee University, South Korea
(E-mail: tskim@khu.ac.kr)

Abstract: In our previous work, a new system using a stereo camera has been proposed to estimate human motion from a sequence of 3-D video frames [1]. The system used the articulated human model defined with connected ellipsoids and then co-registered the 3-D human model to 3-D data to recover the 3-D human body posture on each frame. Consequently, the 3-D human motion is reflected by the kinematic angles of the estimated human body postures. However, this approach had a limitation of the prolonged computational time, because, in order to fit the 3-D human model to the 3-D data, a large number of 3-D points were used in the co-registration. In this paper, we have proposed a new co-registration algorithm that uses the Gaussian distribution to cluster of a set of 3-D points into groups for registration. This improvement leads to the reduction of computations and allows our algorithm to be able to process about 10 frames per second and to be more suitable for real time applications.

Keywords: Human motion capturing, 3-D human model, stereo vision.

1. INTRODUCTION

In recent years, a lot of systems using computer vision techniques to capture human motion have been developed for many applications that involve the interactions between humans and electronic devices such as the remote control using body gestures, games, and robotics. Among utilized technologies, a system using a stereo camera for human motion capturing is now receiving more attention, since the stereo camera is comfortable to users and robust to work on a wide range of applications. Here, the stereo camera refers to the camera that is capable of synchronously acquiring two images of the scenes in slightly different view angles to estimate the 3-D information of the world around. Prior to the use of the stereo camera, the conventional approaches for 3-D human motion recovery were common with the markers and multiple cameras. However, the marker-based approach is inconvenient to users in daily life activities. Meanwhile, the multiple camera approach [2, 3, 4] requires a fixed installation, consequently limiting the flexibility of a system using multiple cameras. In addition, both systems need expensive equipment, especially with the marker-based approach. Thus, an approach using an inexpensive and flexible camera like the stereo camera is more preferred to these approaches. Obviously, reconstructing a 3-D human body posture from the 3-D data recorded by the stereo camera is a difficult problem because it is related to estimate highly kinematic parameters of an articulated human model from the 3-D data. In [5], Yang et al. proposed an approach to tackle this problem based on the matching based approach: The matching based approach uses the comparison mechanism to find a human posture from a database best matching with the 3-D information obtained by the stereo camera. However, there might be a huge number of possible human postures. Therefore, this approach needs a large database to store all postures and an efficient searching method to retrieve postures from the database.

In the earlier work, a system to estimate the human motion from a sequence of 3-D data captured by the stereo camera [1] was developed by us. This algorithm was built from the model-based approach to avoid maintaining a database of human poses. In addition, the algorithm utilized both 3-D data and 2-D cue images for better registration and was more generic to integrate any useful information of the body part detection to improve the accuracy of the human posture recovery. The testing results of the algorithm on the real and synthetic data has shown that the algorithm was able to directly estimate the human motion from the depth images even for complicated human movements without the need of using markers. However, the proposed system has a limitation with somewhat prolonged computational time. The registration between a large number of 3-D points and the kinematic parameters causes the algorithm to take too many computations to recover a right human posture in each frame.

In this work, we develop an improved registration algorithm between the 3-D human model and the 3-D data by grouping the points belonging to a same body part into a Gaussian distribution. Thus, the registration process changes the computations from thousands of points to a few specific parameters of the Gaussian distributions, consequently reducing the computational time. Meanwhile, our algorithm does not affect much to the accuracy, as being shown by the experimental results in Section 3.

This paper is organized as follows. In Section 2, we introduce about the mathematics and algorithm of our approach. The experimental results are presented in Section 3. Finally, we conclude with some discussion in Section 4.

2. METHODS

We utilize the stereo camera to capture the 3-D information. The 3-D information is typically reflected in a 2-D image called a depth image in which depth information is encoded in grayscale values. We extract the region of interest of the depth image correspondent to a tracked person, by considering the distance from the

tracked person to the stereo camera. The depth image is then sampled by a grid to extract a set of points in order to convert the depth image into 3-D data. An example of a depth image and the 3-D data computed from the depth image is depicted in Fig. 1.

The 3-D human model is created by a set of ellipsoids that are connected by kinematic chains and specified by the kinematic parameters. In this work, we mainly focus on the upper body tracking, so we utilize 6 ellipsoids to model the 3-D human body. There are three degrees of freedom (DOF) at each body joint and three translations from the global coordinate system to the local coordinate system at the human hip. In total, we have 21 DOF of the kinematic parameters of the 3-D human model.

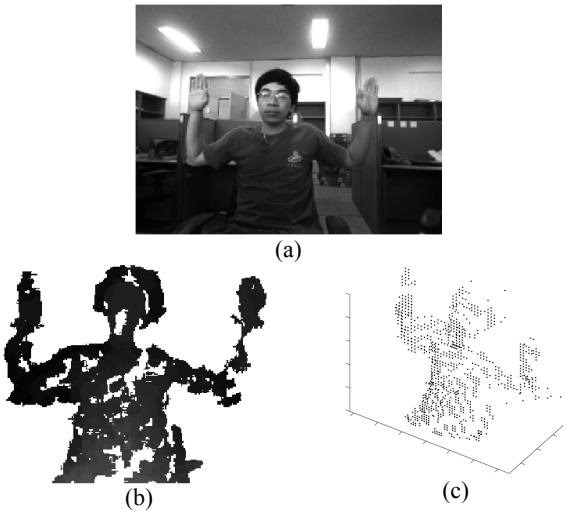

Fig. 1 Data captured by the stereo camera. (a) RGB image. (b) The region of interest of a depth image corresponding to a human subject. (c) 3-D data.

The co-registration algorithm between the 3-D data and the 3-D human model is performed by an iterative procedure, described in Fig 2.

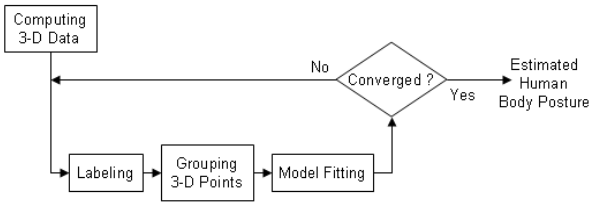

Fig. 2 An overview of our proposed system.

2.1 Labeling and Grouping 3-D Points

Each point of the 3-D data is assigned one label, corresponding to one part of the human body. For simplification, the face detection, skin detection, and geodesic constraint presented in [1] are eliminated when we perform the label assignment. The only criterion utilized for the label assignment is the Euclidean distance from one point to the ellipsoids of the human model. Each point will get a label correspondent to the nearest ellipsoid. In the new approach, a group of pixels with the same label will be represented by the Gaussian distribution. From a set of points $X_j \subset$ body part *i*, the raw Gaussian distribution is estimated by $P(x_i \mid D) \propto \mathcal{N}(m_{i0}, \Sigma_{i0}^{-1})$, where the mean m_{i0} is $\frac{1}{N_i}\sum_j X_j$ and the covariance matrix Σ_{i0}^{-1} is $\frac{1}{N_i - 1}\sum_j (X_j - m_{i0})(X_j - m_{i0})^T$ (N_i is the number of pixels belonging to the body part *i*). Since each Gaussian distribution represents one part of the human body, its shape must be proportional to the size of the corresponding ellipsoid. Thus, the correct estimated Gaussian distribution $\tilde{P}(x_i) \propto \mathcal{N}(m_i, \Sigma_i^{-1})$ must have the covariance matrix Σ_i^{-1} in the form of $R_i^T \Lambda_{i0} R_i$, in which $\Lambda_{i0} = diag\{a^{-2}, b^{-2}, c^{-2}\}$ is a constant matrix, formulated the shape of the Gaussian distribution. Three constants *a, b,* and *c* are the sizes of the ellipsoid along the *x-, y-,* and *z-* axis and R_i is the rotation matrix of the Gaussian distribution, satisfying the constraint $R_i^T R_i = I$. Our main target is to estimate the Gaussian distribution $\tilde{P}(x_i)$ as close as possible to $P(x_i \mid D)$, so we have to minimize the Kullback-Leibler (KL) distance between them

$$\min_{R_i, m_i} J_i(R_i, m_i) = \int \tilde{P}(x_i) \log \frac{\tilde{P}(x_i)}{P(x_i \mid D)} dx_i \quad (1)$$

subject to $R_i^T R_i = I$

or equivalent to

$$\min_{R_i, m_i} J_i(R_i, m_i) = \int \tilde{P}(x_i) \log \tilde{P}(x_i) dx_i - \int \tilde{P}(x_i) \log P(x_i \mid D) dx_i \quad (2)$$

subject to $R_i^T R_i = I$.

We use $\Sigma_i^{-1} = R_i^T \Lambda_{i0} R_i$ and $\tilde{P}(x_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x_i - m_i)^T \Sigma_i^{-1} (x_i - m_i)\right)$ to rewrite the equation (2) as follows,

$$\min_{R_i, m_i} \Im_i = -\frac{1}{2}\log|\Sigma_i| + \frac{1}{2}(m_i - m_{i0})^T \Sigma_{i0}^{-1} (m_i - m_{i0}) + \frac{1}{2} Tr\left(\Sigma_{i0}^{-1} \Sigma_i\right) \quad (3)$$

subject to $R_i^T R_i = I$.

In addition, two connected body parts *(i,j)*, e.g., the upper hand and lower hand, require a constraint to draw them close together,

$$\left\| m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji} \right\|^2 = 0 \quad (4)$$

where the two vector a_{ij} and b_{ji} are described in Fig 3.

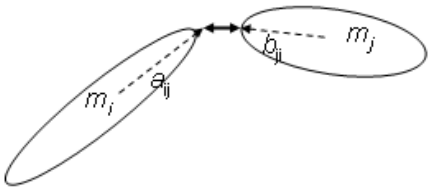

Fig. 3 The distance constraint between two connected parts.

The overall optimization formulation to compute the Gaussian parameters of each body part is rewritten by

$$\min_{R_i, m_i} \Im_i \quad \text{subject to } R_i^T R_i = I, \quad \left\| m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji} \right\|^2 = 0, \forall j \in S_i \tag{5}$$

where S_i is a set of body parts connected to the body part *i*. The complicated optimal problem in equation (5) with equality constraints can be solved by the gradient descent method using the additional Lagrange multipliers [6]. We add a set of Lagrange multipliers λ_{ij} into the optimization problem in (3) to define a new Lagrange function,

$$\Lambda_i(m_i, R_i, \lambda_{ij}) = -\frac{1}{2}\log|\Sigma_i| + \frac{1}{2}(m_i - m_{i0})^T \Sigma_{i0}^{-1}(m_i - m_{i0} + \frac{1}{2} Tr\left(\Sigma_{i0}^{-1}\Sigma_i\right) + \sum_{j \subset S_i} \lambda_{ij} \left\| m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji} \right\|^2. \tag{6}$$

Here, it is not necessary to concern the Lagrange multipliers for the equality constraint $R_i^T R_i = I$, since this constraint can be satisfied by using the symmetric orthonormalization method [7]: Whenever the value of R_i is changed, we normalize R_i by an equation

$$R_i \leftarrow R_i (R_i^T R_i)^{-1/2}. \tag{7}$$

In order to derive the update rules to compute the values of m_i and R_i by the gradient descent method, we need to compute the derivatives of Λ_i with respect to m_i and R_i,

$$\frac{\partial \Lambda_i}{\partial m_i} = (m_i - m_{i0})^T \Sigma_{i0}^{-1} + 2\sum_{j \subset S_i} \lambda_{ij}(m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji})^T$$
$$\frac{\partial \Lambda_i}{\partial R_i} = -R_i + R_i \Sigma_i \Sigma_{i0}^{-1} + 2\sum_{j \subset S_i} \lambda_{ij} a_{ij}(m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji})^T. \tag{8}$$

Finally, a set of equations used to update the values of m_i, R_i, and λ_{ij} until convergence are given by [6],

$$\lambda_{ij} \leftarrow \lambda_{ij} + \eta \left\| m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji} \right\|^2$$
$$m_i \leftarrow m_i - \gamma \begin{pmatrix} m_i - m_{i0} \\ + 2\sum_{j \subset S_i} \lambda_{ij}(m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji}) \end{pmatrix}$$
$$R_i \leftarrow R_i + \gamma \begin{pmatrix} R_i - R_i \Sigma_i \Sigma_{i0}^{-1} \\ -2\sum_{j \subset S_i} \lambda_{ij} a_{ij}(m_i + R_i^T a_{ij} - m_j - R_j^T b_{ji})^T \end{pmatrix} \tag{9}$$
$$R_i \leftarrow R_i (R_i^T R_i)^{-1/2}.$$

Where η and γ are learning rates. Here, in order to move m_i to its raw estimation m_{i0} faster, we replace $\Sigma_{i0}^{-1}(m_i - m_{i0})$ by $(m_i - m_{i0})$ in the computation of m_i.

2.2 Model Fitting

After the parameter (R_i, m_i) of each Gaussian distribution is found from the previous stage, the kinematic parameters $\theta_1, \theta_2, ..., \theta_k$ are calculated from a set of equations

$$R_x(\theta_1)R_y(\theta_2)R_z(\theta_3)R_x(\theta_4)...R_y(\theta_{k-1})R_z(\theta_k) = R_i \tag{10}$$

where $R_x(\theta_n), R_y(\theta_n)$, and $R_z(\theta_n)$ are the rotation matrices around the *x*-,*y*-, and *z*- axis by an angle θ_n.

For simplification, in our algorithm, we compute the kinematic parameters for two body parts at each time, e.g., head and body or left lower hand and left upper hand. For the two connected parts *a* and *b*, we have a couple of equations to find the kinematic parameters

$$\begin{cases} R_x(\theta_1)R_y(\theta_2)R_z(\theta_3) = R_a \\ R_x(\theta_1)R_y(\theta_2)R_z(\theta_3)R_x(\theta_4)R_y(\theta_5)R_z(\theta_6) = R_b \end{cases} \tag{11}$$

where $R_a = [r_a^{i,j}]_{i=1,2,3,j=1,2,3}$ and $R_b = [r_b^{i,j}]_{i=1,2,3,j=1,2,3}$ are the two rotation matrices of the two parts.

The values of θ_1, θ_2, and θ_3 are computed from R_a by

$$\theta_2 = \arcsin(r_a^{3,1})$$
$$\theta_1 = \arctan 2\left(r_a^{3,2}/\cos(\theta_2), r_a^{3,3}/\cos(\theta_2)\right) \tag{12}$$
$$\theta_3 = \arctan 2\left(r_a^{2,1}/\cos(\theta_2), r_a^{1,1}/\cos(\theta_2)\right)$$

where $\arctan 2(x, y)$ is a function that returns a value of $\arctan(x/y)$ or $\arctan(x/y) + \pi$ depending on the quadrant in which the point *(x,y)* belongs to. Note that for special cases $r_a^{3,1} = \pm 1$, we need new equations to estimate the kinematic angles:

If $r_a^{3,1} = 1$ then we have

$$\theta_1 = \arctan 2(r_a^{1,2}, r_a^{1,3}), \theta_2 = -\pi/2, \theta_3 = 0; \tag{13}$$

And if $r_a^{3,1} = -1$ then we have

$$\theta_1 = \arctan 2(r_a^{1,2}, r_a^{1,3}), \theta_2 = \pi/2, \theta_3 = 0.$$

The rest of kinematic parameters θ_4, θ_5, and θ_6 are calculated by a similar way using an equation,

$$R_x(\theta_4)R_y(\theta_5)R_z(\theta_6) = R_a^{-1}R_b. \tag{14}$$

3. EXPERIMENTAL RESULTS

We used the Bumblebee stereo camera of Point Grey Research to capture stereo image pairs with the resolution of 320×240. The built-in algorithm attached along with the stereo camera was applied to estimate the 3-D data from the image pairs. The subject performed activities about 0.5–1.5 meters from the camera. We compare the computational time of our algorithm to estimate the human body postures from the 3-D data with the previous approach [1]. The comparison results given in Table 1 have shown a significant improvement of our algorithm in the processing time aspect, compared with the previous one. Besides, to qualitatively evaluate the recovering accuracy of our approach with the previous one, we show the results of the estimated upper body postures for some particular frames in Fig. 4. Obviously, the two methods get similar accuracy of recovering the 3-D human postures from stereo data.

Table 1 The comparison between our new approach and the previous approach [1] in the computational time.

	Previous approach	New approach using the Gaussian distribution to group 3-D points
Total processing time for 200 frames (s)	37.34	19.73
Frames per second	5.36	10.14

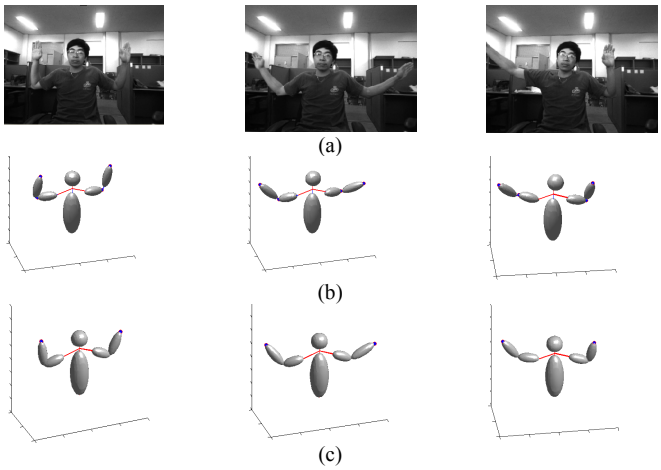

Fig. 4 (a) The right images of stereo image pairs and the recovered human postures of the image sequences estimated by our approach (b) and the previous one (c).

4. CONCLUSION AND DISCUSSION

In this paper, we have proposed a new algorithm to improve the process of fitting an articulated human model to the 3-D data. Rather than directly calculating on a huge number of points, we have grouped the points with a same label into a group parameterized by the Gaussian distribution. Consequently, we have reduced the computations of the registration process that leads to a remarkable improvement of our algorithm in the processing rate to about 10 frames per second.

In future work, in addition to using the Euclidean distance for the label assignments as presented in this paper, we plan to utilize more information from 3-D data and RGB images as used in the previous work [1] to improve the labeling step of our co-registration algorithm that might consequently increase the accurate performance of our human motion capturing system.

ACKNOWLEDGMENT

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2009-(C1090-0902-0002)). This work also was supported by a Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST No. 2008-1342) and was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (2009-0076798).

REFERENCES

[1] N. D. Thang, T. -S. Kim, Y. -K. Lee, and S. Lee, "Estimation of 3-D human body posture via co-registration of 3-D human model and sequential stereo information," *Applied Intelligence*, *in press*, (2010).

[2] A. Sundaresan and R. Chellappa, "Model driven segmentation of articulating humans in Laplacian Eigenspace", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 10, 1171-1785, (2008).

[3] R. Horaud, M. Niskanen, G. Dewaele, and E. Boyer, "Human motion tracking by registering an articulated surface to 3D points and normals", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 1, 158-163 (2009).

[4] D. Knossow, R. Ronfard, and R. Horaud, "Human motion tracking with a kinematic parameterization of extremal contours", *International Journal of Computer Vision*, Vol. 79, No. 3, 247-269, (2008).

[5] H. –D. Yang and S. –W. Lee, Reconstruction of 3D human body pose from stereo image sequences based on top-down learning, *Pattern Recognition*, Vol. 40, No. 11, 3120-3131, (2007).

[6] W. Lu and J. C. Rajapakse, "Approach and applications of constrained ICA", *IEEE Trans. on Neural Networks*, Vol. 16, No. 1, pp. 203-212, (2005).

[7] A. Hyvärinen, J. Karhunen, and E. Oja, "*Independent component analysis*", John Wiley and Sons, 2001.