Vision-Based Motion Capture of Interacting Multiple People

Hiroaki Egashira[1], Atsushi Shimada[1], Daisaku Arita[1,2], and Rin-ichiro Taniguchi[1]

[1] Department of Intelligent Systems Kyushu University
744, Motooka, Nishi-ku, Fukuoka, 819-0395, Japan
{aki,atsushi,rin}@limu.is.kyushu-u.ac.jp
[2] Institute of Systems, Information Technologies and Nanotechnologies
2-1-22, Momochihama, Sawara-ku, Fukuoka, 814-0001, Japan
arita@isit.or.jp

Abstract. Vision-based motion capture is getting popular for acquiring human motion information in various interactive applications. To enlarge its applicability, we have been developing a vision-based motion capture system which can estimate the postures of multiple people simultaneously using multiview image analysis. Our approach is divided into the following two phases: at first, extraction, or segmentation, of each person in input multiview images; then, posture analysis for one person is applied to the segmented region of each person. The segmentation is realized in the voxel space, which is reconstructed by visual cone intersection of multiview silhouettes. Here, a graph cut algorithm is employed to achieve optimal segmentation. Posture analysis is based on a model-based approach, where a skeleton model of human figure is matched with the multiview silhouettes based on a particle filter and physical constraints on human body movement. Several experimental studies show that the proposed method acquires human postures of multiple people correctly and efficiently even when they touch each otter.

1 Introduction

Motion capture systems (MCS) are useful tools for various interactive applications such as interactive animation, 3D virtual space operation, video game interface, etc. Especially, vision-based MCS is a smart and natural approach since it does not impose any physical restrictions on a user.

There are many researches into vision-based MCS: some of them employed a computation intensive approach to acquire precise motion information [1][2][3] and others made emphasis on real-time features so that it can be applied to interactive applications [4][5][6][7]. However, most of them handle posture estimation of just one person. From the viewpoint of human activity observation, it is also required to acquire motion information of multiple people who are interacting with each other.

To analyze the motion of multiple people, Tanaka et al. [8] employed an example-based approach. They made a database which consists of combined 3D

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 451–460, 2009.

shapes of two persons and their posture parameters, and estimated the postures by searching the database for observed 3D shapes. However, when multiple people are interacting with each other, the variation of the 3D shapes of the people becomes quite large and it is not easy to efficiently find a solution in such a large search space. This difficulty becomes larger when the number of persons becomes large. In contrast, we have employed another approach, where each person in the scene is segmented, and where the posture of the segmented person is analyzed by an ordinary one-person motion capture system. In this paper, we will present our motion capturing of multiple persons, emphasizing on the segmentation of the multiple persons using multiview image analysis. Also, we will show several experimental results showing the effectiveness of our algorithm.

In general, there are two approaches to segment human regions; one is segmentation in the 2D image space and the other is segmentation in the 3D space, where the 3D shape of the target object is reconstructed from multiview images. We have employed the latter approach, and the main issue is to assign a correct person identification label to each voxel, which represents the 3D shape of the reconstructed object. In this paper, we assume that there are two people in the observed space. When a person does not touch the other, it is easy to assign a unique label to each human region by finding a connected component of voxels [9,10]. However, when two people are interacting with each other, they often touch each other and their reconstructed 3D shape becomes one connected component. In this case, we have to segment the two persons. To solve this problem, we define an energy function that expresses the identifiability of the persons, which is assigned to each voxel of the reconstructed shape, and we realize the segmentation by minimizing the total energy assigned to segmented regions based on a graph cut algorithm [11].

2 Human Region Segmentation in the 3D Space

2.1 Outline

The goal of this human object segmentation is to extract regions each of which represents one person. In general, the segmentation in the 2D image is rather difficult because when we observe multiple persons, especially multiple persons interacting with each other, they usually make serious self-occlusion in the 2D image space. Here, since target people are observed by multiple cameras, we can construct their 3D shape from the acquired images, and it is much easier to classify the target region into separate human regions in the 3D space, each of which represents one person, although the computation time becomes larger due to the reconstruction process.

According to the above consideration, first, we reconstruct the 3D shape of the target from multiview images based on visual cone intersection [13], where the 3D shape is represented in terms of voxels. Then, we segment the reconstructed 3D target shape into separate 3D regions of the observed people. Visual cone intersection does not generate a precise 3D shape of the object but the purpose of the reconstruction here is just to acquire cues for multiple person segmentation.

From this point of view, we do not have to make the resolution of the voxel space high, and, thus, the computation cost is not increased heavily. After each person is segmented, a vision-based MCS for one person [7] is applied to the segmented data.

To partition the reconstructed 3D shape, we have employed an energy minimization framework. We attach a label, i.e., a person identifier, to each voxel, and we define an energy function that expresses the suitability of the person identifier attached to the voxel. To achieve its fast computation, we have segmented human regions by a graph cut algorithm [11,12] minimizing the total energy of all the voxels belonging to segmented regions. In general, the energy function can be defined based on the following information:

Temporal sequence information: the segmentation results in the previous frames are used to evaluate the validity of the current segmentation,

Visual features: 2D/3D visual features extracted from the images are referred to.

In the former, the difference of the voxels in the current frame and those in the previous frame is calculated, and each voxel of the detected difference is categorized according to its distance to the segmented regions in the previous frame. On the contrary, in the latter, each voxel is categorized according to its distance to detected robust visual features, i.e., skin-color blobs (shoe-color blobs for feet), of each person. Although the former is quite effective to reduce the computation cost and to realize real-time processing, it requires an initial segmentation and, more seriously, it tends to accumulate the segmentation error as the segmentation process proceeds into the succeeding frames. Therefore, here, we have not used the previous segmentation result, and we have segmented human regions based on the visual features extracted by image analysis.

2.2 Energy Minimization

The energy function we define consists of the data term and the smoothing term as follows.

$$E(X) = kG(X) + H(X) \tag{1}$$

$X = (X_1, \ldots, X_v, \ldots, X_{|V|})$ is a binary vector, where X_v is a label of a voxel v: person A or B[1], and V is a set of voxels to be labeled. $k(> 0)$ is the ratio of the data term $G(X)$ and the smoothing term $H(X)$. $G(X)$ is defined, referring to $g(X_v)$, the suitability of a given label, as follows:

$$G(X) = \sum_{v \in V} g(X_v) \tag{2}$$

$$g(X_v) = \begin{cases} g_A(v), & \text{if } X_v = A \\ g_B(v), & \text{if } X_v = B \end{cases} \tag{3}$$

[1] In this paper, we assume two persons are observed. However, the idea can be extended for analyzing three or more people.

where $g(X_v)$ denotes the likelihood of a given label. The detailed explanation will be given in the following subsection. On the other hand, $H(X)$ is defined, referring to $h(X_u, X_v)$, the consistency between neighboring pixels, as follows:

$$H(X) = \sum_{(u,v)\in N} h(X_u, X_v) \tag{4}$$

$$h(X_u, X_v) = \begin{cases} 1, & \text{if } X_u \neq X_v \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

N is a set of all the neighboring pixel pairs in V here. We assign proper labels to the voxels which minimize the total energy $E(X)$, and it is solved by a graph cut algorithm.

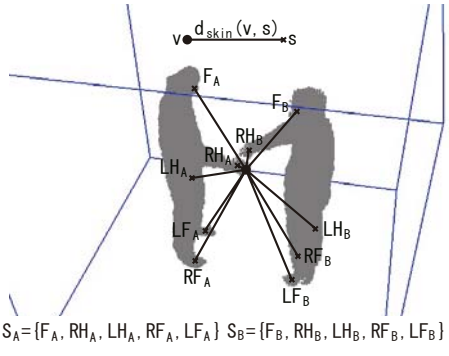

Fig. 1. 3D Positions of the Color Blobs

2.3 Human Region Segmentation Using Color Blob Information

Skin color blobs are good visual features for detecting people, and we have developed a segmentation method based on skin color blobs. We classify the voxels of reconstructed object shape, V, into person A or person B based on the distance to the color blobs of person A and those of person B. Blob identification is accomplished based on the blob positions in the previous frame, and basically the label of the nearest blob which has similar features in the previous frame is assigned to a given blob[2].

We obtain 3D positions of skin-color blobs (shoe-color blobs for feet) by multiview image analysis. Here, let a color blob be s, and a set of color blobs which belong to person A be S_A, and S_B for person B. A person has five color blobs; face (F), right hand (RH), left hand (LH), right foot (RF) and left foot (LF). Then, we calculate $g_A(v)$ as follows (see Fig.1).

[2] We assume, at beginning, the two persons do not touch each other and color blobs are easily classified into person A or B based on connected component analysis.

$$g_A(v) = \sum_{s \in S_B} \frac{1}{d_{skin}(v, s)} \tag{6}$$

$$S_B = \{F_B, RH_B, LH_B, RF_B, LF_B\} \tag{7}$$

$$d_{skin}(v, s) = D_E(P_V(v), P_S(s)) \tag{8}$$

Here, $D_E(\cdot, \cdot)$ denotes a Euclidean distance between the two points, $P_V(\cdot)$ the 3D position of the voxel, and $P_S(\cdot)$ is the 3D position of the color blob. If a voxel v is far from the color blobs of the other person, the value of $g_A(v)$ becomes small. We expect that segmentation results become good by considering all the color blobs belonging to the other person.

3 Human Posture Analysis

After the result of human region segmentation in the 3D space is projected to the multiview image space, we apply a one-person pose estimation method to each segmented human region. In our human posture analysis, to estimate full body human postures, we have introduced a skeleton model with 33 DOF (3DOF translation and 30 DOF orientation, see Fig.2). This model includes 4 DOF on the torso part, which enables the user to move flexibly, such as bending or leaning poses. We have established the human figure model relatively complex to make generated human postures natural.

Our algorithm is skeleton-based model fitting, in which the skeleton model of the human body is fit to the center of a human silhouette. In general, the model fitting approach for motion capturing iteratively synthesizes human models and analyzes their fitness based on image features, and it is usually time consuming because the human configuration space is very high dimensional. To solve this problem, we also use 3-D positional constraints among the body parts. For example, when we estimate the posture of the left arm, we can do much faster if the 3-D position of the left hand is known. The basic algorithm flow of our real-time motion capture system is as follows:

1. Detection of visual cues
   - Silhouette detection and skin-color blob (and shoe-color blob) detection.
   - Calculation of the 3-D positions of the color blobs using multiview fusion. The 3-D head position is also precisely estimated by Hough transform, which searches for a circular silhouette edge around the detected skin-color blobs.
   - Generating the distance-transformed image of the silhouette region for fast skeleton fitting in the next step.
2. Estimation of human posture
   Fitting the skeleton model of the human body to the center of silhouette using particle filter. Several constraints to narrow the search space have been introduced as follows:

constraints based on collision detection. The torso and the limbs cannot occupy the same place in 3-D space at the same time. We eliminate postures which have collisions among the torso and the limbs.

kinematic constraints. Human joints have movable limits. For example, the elbows cannot bend backward. We establish the limits on each joint angle based on anatomy to eliminate impossible postures and to prune the model configuration space.

Fig.3 shows the framework of our motion sensing. In the stage of posture estimation, first, we estimate the posture of the torso part. Next, four limb postures are estimated separately by using neck or waist position calculated in the previous step.

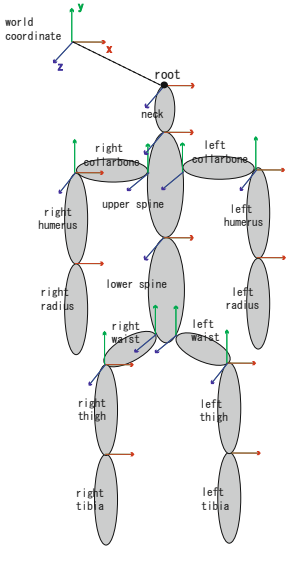

Fig. 2. Human Figure Model

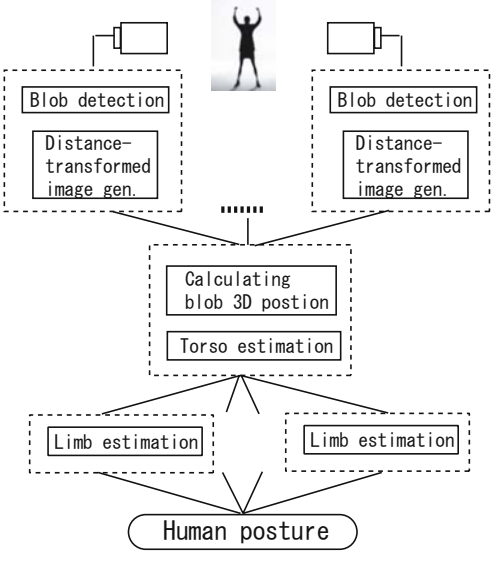

Fig. 3. System Overview

4 Experimental Results

4.1 Human Region Segmentation

We have examined our algorithm using the following three video sequences which include scenes where a person touches the other:

Sequence1: person A shakes person B's hand (16frames).
Sequence2: person A touches, by his right hand, the left shoulder of person B (12frames).
Sequence3: person A kicks, by his right foot, person B on the left thigh (20frames).

In this experiment, we used 8 cameras (Dragonfly 2 of Point Grey Research) which had been calibrated in advance [14]. The sizes of captured images are 640×480 pixels and the resolution of the voxel space is $64 \times 64 \times 64$. We obtained segmentation results shown in Fig.4, 5 and 6, where typical three cases are shown.

Before touching: two people have not touched yet (the upper rows).
Just touching: two people are just touching (the middle row).
After touching: two people become untouched (the lower rows).

Segmentation referring to the 3D positions of the color blobs mostly succeeded, except the case that blobs of different persons are merged into one. In this case, we have to attach dual labels to the blob, by which the blob in the 3D space is projected onto both of 2D image regions corresponding the two persons.

The computational time for this segmentation is about 9ms/frame with Pentium4 3GHz CPU. Although the computational time to obtain the color blobs is not included, the color blob detection is also required in the human posture estimation phase, and it doe not become a large overhead of the system.

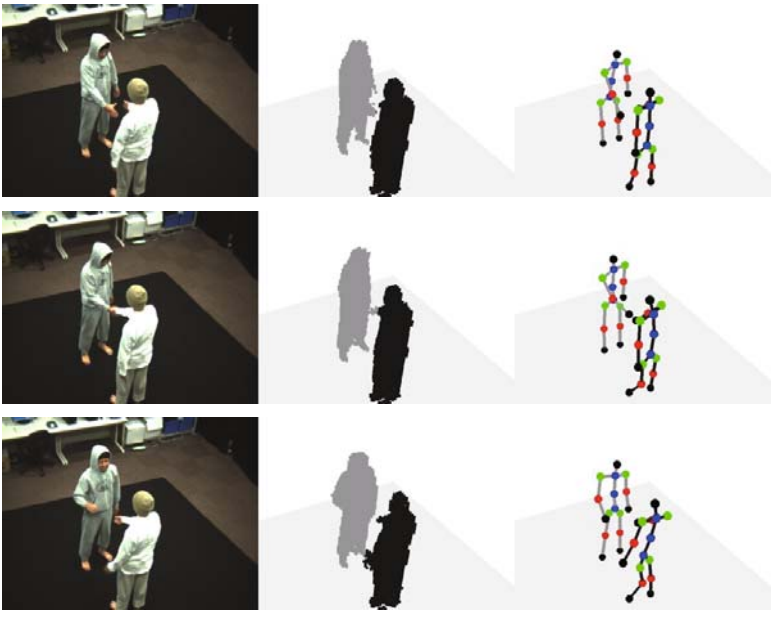

Fig. 4. Segmentation Results of Sequence1. Left: Input images, Middle: Segmentation results. Right: Estimated human postures.

4.2 Results of Human Motion Capturing

The right columns of Fig.4, 5, 6 show the results of human posture analysis of multiple people. They show that the postures of the people are correctly analyzed

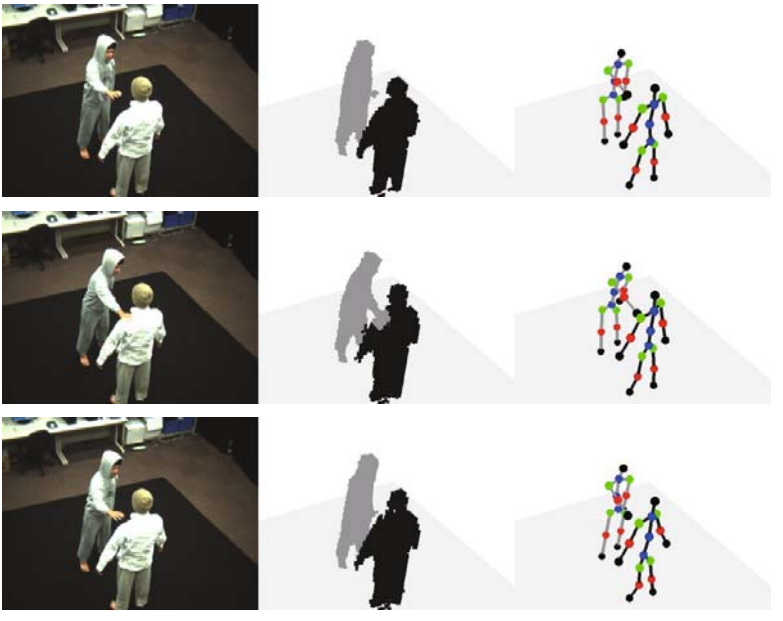

Fig. 5. Segmentation Results of Sequence2

even when they touch each other. The computation cost is summarized in Table 1 (Pentium4 3GHz CPU), which shows the algorithm can be executed in real-time when we implement it on a parallel machine such as a PC-cluster [15].

Table 1. Computation Time of Posture Estimation

Category	Comp. Time
Visual Feature Extraction	
–Skin Blob Extraction	4msec
–Distance Image Generation	60msec
–3-D Position Estimation	21msec
Torso Posture Estimation	49 msec
Arm Posture Estimation	95 msec
Leg Posture Estimation	129 msec

Our human region segmentation relies on the blob detection, and therefore, when the blobs are not correctly detected, the result of human posture analysis is seriously affected. To apply our method to uncontrolled, i.e., complex background, we need robust color blob detection, and we have employed example-based color detection.

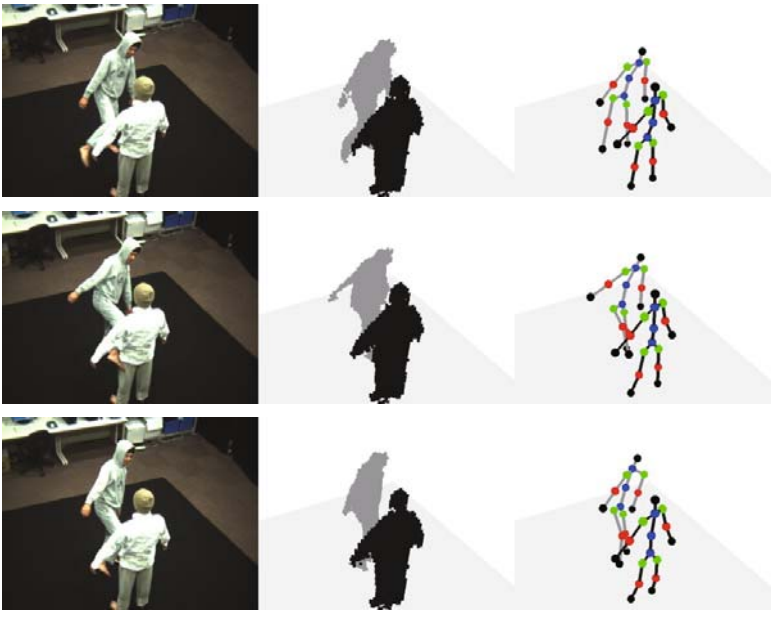

Fig. 6. Segmentation Results of Sequence3

5 Conclusion

In this paper, we have proposed a motion capture system to estimate the postures of interacting multiple people. To achieve the goal, we have constructed a two phase system, which consists of region segmentation of each person and human posture estimation of each segmented person. The segmentation is realized in the voxel space and a graph cut algorithm is employed to achieve optimal segmentation. Posture analysis is based on a model-based approach, where a skeleton model of human figure is matched with the multiview silhouettes. Experimental results have indicated the effectiveness of our method, showing that it acquires human postures of multiple people correctly and efficiently even when they touch each other.

In our future work, we are going to thoroughly evaluate the accuracy of the system, and to make the system more robust so that the system can be used in more practical environment with complex background. Also observing interacting people with tools or objects is the next goal.

References

1. Sundaresan, A., Chellappa, R.: Multi-camera Tracking of Articulated Human Motion Using Motion and Shape Cues. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3852, pp. 131–140. Springer, Heidelberg (2006)

2. Carranza, J., Theobalt, C., Magnor, M., Seidel, H.: Free-Viewpoint Video of Human Actors. In: Proc. of ACM SIGGRAPH, pp. 569–577 (2003)
3. Sand, P., McMillan, L., Popovic, J.: Continuous Capture of Skin Deformation. In: Proc. of ACM SIGGRAPH, pp. 578–586 (2003)
4. Date, N., Yoshimoto, H., Arita, D., Taniguchi, R.: Real-time Human Motion Sensing based on Vision-based Inverse Kinematics for Interactive Applications. In: Proc. of International Conference on Pattern Recognition, vol. 3, pp. 318–321 (2004)
5. Kehl, R., Bray, M., Van Gool, L.: Full Body Tracking from Multiple Views Using Stochastic Sampling. In: Proc. of Computer Vision and Pattern Recognition, pp. 129–136 (2005)
6. Bernier, O.: Real-Time 3D Articulated Pose Tracking using Particle Filters Interacting through Belief Propagation. In: Proc. of International Conference on Pattern Recognition, vol. 1, pp. 90–93 (2006)
7. Saiki, T., Shimada, A., Arita, D., Taniguchi, R.: A Vision-based Real-time Motion Capture System using Fast Model Fitting. In: CD-ROM Proc. of 14th Korea-Japan Joint Workshop on Frontiers of Computer Vision (2008)
8. Tanaka, H., Nakazawa, A., Takemura, H.: Human Pose Estimation from Volume Data and Topological Graph Database. In: Proc. of 8th Asian Conference on Computer Vision, pp. 618–627 (2007)
9. Sagawa, Y., Shimosaka, M., Mori, T., Sato, T.: Fast Online Human Pose Estimation via 3D Voxel Data. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1034–1040 (2007)
10. Huang, K.S., Trivedi, M.M.: 3D Shape Context Based Gesture Analysis Integrated with Tracking using Omni Video Array. In: Proceedings of IEEE Workshop on Vision for Human-Computer Interaction, V4HCI (2005)
11. Boykov, Y., Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(9), 1124–1137 (2004)
12. Kolmogorov, V.: `http://www.cs.ucl.ac.uk/staff/V.Kolmogorov/software.html`
13. Martin, W.N., Aggarwal, J.K.: Volumetric Description of Objects from Multiple Views. IEEE Trans. on Pattern Analysis and Machine Intelligence 5(2), 150–158 (1983)
14. Tsai, R.Y.: A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. IEEE Journal of Robotics and Automation 3(4), 323–344 (1987)
15. Arita, D., Taniguchi, R.: RPV-II: A Stream-Based Real-Time Parallel Vision System and Its Application to Real-Time Volume Reconstruction. In: Schiele, B., Sagerer, G. (eds.) ICVS 2001. LNCS, vol. 2095, pp. 174–189. Springer, Heidelberg (2001)