2010 International Conference on Pattern Recognition

Human Pose Estimation for Multiple Persons Based on Volume Reconstruction

Xinghan Luo, Berend Berendsen, Robby T. Tan, Remco C. Veltkamp
Department of Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands
xinghan, robby, remco.veltkamp@cs.uu.nl

***Abstract*—Most of the development of pose recognition focused on a single person. However, many applications of computer vision essentially require the estimation of multiple people. Hence, in this paper, we address the problems of estimating poses of multiple persons using volumes estimated from multiple cameras. One of the main issues that causes the multiple person from multiple cameras to be problematic is the present of 'ghost' volumes. This problem arises when the projections of two different silhouettes of two different persons onto the 3D world overlap in a place where in fact there is no person in it. To solve this problem, we first introduce a novel principal axis-based framework to estimate the 3D ground plane positions of multiple people, and then use the position cues to label the multi-person volumes (voxels), while considering the voxel connectivity. Having labeled the voxels, we fit the volume of each person with a body model, and determine the pose of the person based on the model. The results on real videos demonstrate the accuracy and efficiency of our approach.**

***Keywords*-tracking; principal axis; volume reconstruction; pose estimation;**

I. Introduction

Human body model fitting into volumetric (voxel) data has been reported as an efficient and robust approach to recognize single person's poses [1]. Existing methods in this approach [1], [2], [3], [4] use a pre-defined human body model describing the shape, size and connection order of the body parts, to be fitted into the reconstructed 3D human shape features. Subsequently, the methods estimate the pose parameters based on the fitted model. Compared with the other methods based on 2D features in 2D images, pose-recognition methods in 3D data relax the problems of occlusions, either inter-person occlusions or self occlusions [2], [4].

Recent comparisons [1] show more robust estimations can be achieved by employing probabilistic frameworks to model and predict human poses. However, the precisions of all the proposed body model fitting methods are still dependent on the quality of the reconstructed volumetric data. Moreover, the methods focus solely on a single person, making the multiple-persons volumetric-data recovery and pose recognition still an open problem, particularly when people occlude each other. In such a case, multiple-person volume reconstruction requires pixel-wise segmentation of the persons' overlapping silhouettes in 2D views to estimate the volume of each person. However the person segmentation in 2D views is still not yet solved. Previously, Iwashita [5] proposed to use level-set technique to solve the segmentation problem. Guan [6] proposed a Bayesian method to build 3D shapes from multi-person silhouettes cues such as the persons' appearances, combined with the persons' occlusion and localization information, but the algorithm efficiency is not sufficient to pose estimation applications.

In this paper, our ultimate goal is to estimate human poses for multiple people from multiple cameras. To achieve the goal, we take the following steps: (i) 3D tracking for multiple persons (Sect.2), (ii) volume reconstruction and voxel labeling (Sect.3), and (iii) pose estimation through body model fitting (Sect.4). In developing the steps, we built algorithms that distinguish our technique from the existing methods. The following are the list of our contributions: (1) In tracking multiple persons from multiple cameras, we developed a robust technique based on the selection of the best visibility views of different cameras, and use a vertical line to identify person's location based on the vertical coordinate of the 3D world; (2) In labeling the voxels of different persons, we introduce a technique based on the distance of voxel to estimated person position in the 3D world and on the inter-voxel connectivity; (3) To our knowledge, our method is the first attempt to integrate these steps for solving the pose estimation problem for multiple persons.

II. Tracking of Multiple Persons

To reconstruct the 3D volume data, we use the method of shape from silhouettes [7]. We use a fixed lookup table to record 3D voxel to image pixel correspondences in order to improve the efficiency of 3D reconstruction. Note that, in our case, we do not need to have a precise 3D reconstruction, since our ultimate goal is to estimate poses.

The 3D reconstruction produces voxels that represent human bodies, yet unfortunately it does not give any direct indication of which voxels belong to which person. To solve

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.876

this, we need to label the voxels. An important clue for voxel labeling is the position of each person in the 3D world. These positions change when the persons move. Therefore, we need to track those positions simultaneously for different persons.

To track the 3D positions, we first identify the 1D position of every person in 2D images, where the 1D position represents the position in the x-coordinate of a 2D image. We call this 1D position 'principal axis' (since usually it is a line, although it is not necessarily a vertical line w.r.t. the 2D image). The identification of the persons from frame-to-frame is accomplished by using appearance-based probabilistic matching based on the KDE of the colors of the persons [8]. Afterwards, the principal axes of the same person from different views are projected onto the 3D ground plane. In our algorithm, only the axes from the cameras of better visibility are used to estimate the person's 3D ground plane location. We call this best visibility views. Using these views, we no longer need to use all views (from different cameras) in the tracking process. Instead, in practice, we use two best visibility views, see in Fig.1 (b). This technique makes the tracking algorithm more robust, since it can avoid the problems when there are significant occlusions for one or two cameras.

Previous principal-axis based methods such as [8] used the line parallel to vertical image column as the axis, which is in fact not equivalent to a line perpendicular to the ground plane in 3D. To solve this problem, we introduce a Vertical Reference Line (VRL) set, consisting of the 2D correspondences of selected 3D lines perpendicular to the ground floor projected on views. The VRL set provides a better vertical-line-based person-location approximation, and yields a more accurate position estimation in 3D.

In each view, regarding to the appearance models we have built, we label the foreground pixels lie on the VRL set to the persons that the pixels most likely belong to. Then, based on the pixel labels along each VRL, the number of pixels labeled to the same person is summed up to get the histograms of the pixels for the same person. Intuitively, the initial estimate of the multi-person's 2D principal axis is the highest peak of the histogram, see in Fig.1 (a). Since the histogram's peak often corresponds to the VRL that passes through one of the person's legs and reach the top of the head. To find a better VRL that is close to the body symmetric axis, we refine the location of the principal axis by analyzing the distribution of pixels in the left and right side of the initial estimation of the 2D principal axes. See the Fig.1 (a). Additionally, the previously estimated positions in time step $t-1$ constraints the position of the current principal axes, the distance between $t-1$ and t time step should not be larger than a certain threshold, otherwise an alarm of false estimation is indicated.

The estimated principal axes of the same person k from different views are projected onto 3D ground plane as person

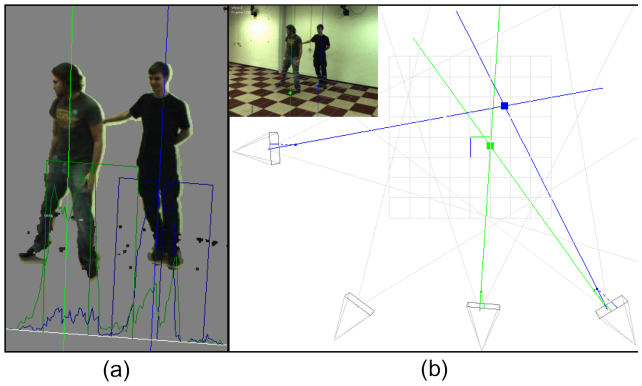

Figure 1. (a) Estimate persons' principal axes in foreground regions. (b) Estimate person 3D ground plane locations (blue and green dots) by best visibility views.

k projection lines, see in Fig.1 (b), blue and green lines. For each time step, the visibility of all the cameras to each person is quantitatively measured by considering two main factors: (1) The distance of the cameras to the person, the shorter the better; (2) The degree of occlusion by the other persons, the lower the better. Based on these criteria, the cameras of best visibility are selected, and used to estimate the person's location in the next time step. Fig.1 (b) shows the views with project lines are selected views of higher visibility to corresponding persons, and the intersections are the estimated 3D location of the persons.

III. Voxel Labeling

Having the constructed unlabeled voxels and the estimated positions of every person in the 3D world, in this section we utilize the positions and the volume-consistency property to label the voxels. The main problem of reconstructing multiple persons' volumes by binary silhouettes without inter-person segmentation is the 'ghost' volume: when the projections of two different silhouettes of two different persons onto the 3D world overlap in a place where in fact there is no person in it. See the light gray regions in Fig.2 (a). Such artifact severely degrades the reconstructed volume. To label voxels and identify the possible 'ghost' volumes, two main labeling criteria are considered: (1) The voxels' distance to the reference locations; (2) Inter-voxel connectivity within the same volume.

A. Labeling By Distance and Connectivity

Given the unlabeled voxels and the 3D positions, we create 3D lines ρ_k ($k \in \{1, ..., N\}$, where N is the number of persons) passing through the persons' estimated positions and perpendicular to the ground plane. Based on each of these 3D lines, we create cylinders with the lines as the center. Regarding the cylinders, we define D_k as the radius of the cylinder of a person k, and can be determined approximately as half of the width of the average of adult's

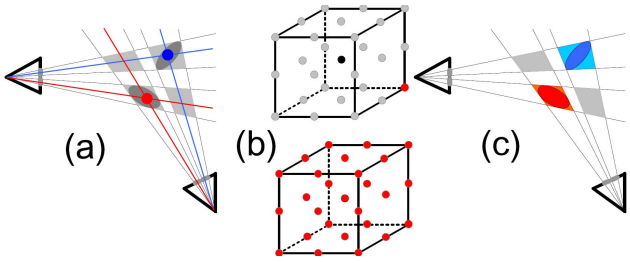

Figure 2. Estimate two persons' volume by location cues and voxel connectivity.

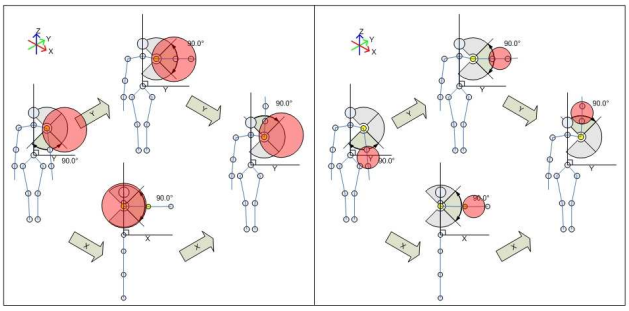

Figure 3. Estimation of upper (left) and lower (right) arm direction.

torsos. Based on the cylinders, we label voxels that intersect with cylinders, and we call them *seed voxels*.

Having the labeled seed voxels, the next step is to label the rest of the voxels. To determine whether the voxels belong to the same person, we consider the inter-voxel connectivity by investigating neighborhood voxels. Similar to [9], we define two voxels $v_1\ (x_1, y_1, z_1)$ and $v_2\ (x_2, y_2, z_2)$ are neighbored if $(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2 \leq 3S$, where S is the unit distance between voxels. See in Fig.2 (b) (top), one 3D voxel (in black) and its 26 neighbors form a cubic volume, any labeled voxel (in red) can infer the rest 26 voxels to have the same label as shown in Fig.2 (b) (bottom).

Following the connectivity constraints, the seed voxel set of each person is expanded while increasing the distance threshold by $2S$ for each time step. For voxel that is within the 3D Euclidean distance $D_k+(G-1)S < D \leq D_k+GS$ to ρ_k (where G is the index of the voxels group), check its 26 neighboring voxels (if available): If there is at least one seed voxel, label the voxel and all its neighbors as seed voxels. If there is no neighbor, or no labeled neighbor, then label the voxel as 'unknown'.

After labeling the voxels according to the above two criteria, the volume of each person is estimated, see the colored regions in Fig.2 (c). The ghost volumes that do not connect to any person's actual volumes are therefore identified in gray regions.

IV. Pose Estimation for Multiple Persons

Having estimated the multi-person ground plane positions and the voxels, the aim of this section is to discuss how we estimate the poses of multiple persons. Our approach is principally based on volume-based body model fitting [2], [3], [4]. First, the human model positions are tracked based on the persons' ground plane positions (Sect.2) and then the constructed voxels are labeled (Sect.3). After the labeling, the human models are to fit to the labeled voxels to provide the pose parameters of the persons.

A. Human Body Model

To extract pose parameters, a simplified generic articulated model for basic shape estimation of a human body is used, see in Fig.3. It consists of a 10-joint skeleton representing the basic human kinematic structure, based on which the torso, arms and legs are modeled as cylinders and the head as a sphere, all are simple volumetric primitives. The human body parts of the model are defined in relation to body length. At the beginning, the persons are required to stand straight in T-shape poses to initialize the body models. The persons' estimated 3D positions determine the initial locations of the body models. According to the estimated 3D heights of the persons, the lengths of the models' body parts are estimated by anthropometric measurements [3].

B. Body Model Fitting

The fitting order follows a hierarchical approach with common sense heuristics: the head of the person is first located, then the neck and pelvis points are found which determine the torso. From the torso the shoulders can be located, followed by the elbows and hands. In relation to the shoulders and pelvis, the hips are located, followed by knees and feet. This hierarchical approach requires multiple or reoccurring generic algorithms for each body model part.

Indirect body model fitting methods [2], [3], [4] are used to fit body model parts to each person's estimated volume. For each fitting step, template body part is placed on certain voxel group to find center of voxel masses as feature point to determine the start and end point of a body part, and voxels inside the template body part are marked as known part of the body, and thus excluded from the next fitting step. See in Fig.3 (left), given the estimated shoulder position as the reference, the upper arm direction is made by placing a sphere next to the shoulder. The centroid of the unmarked voxels within the spherical region is computed. If no voxels are found within the region, then the radius of the sphere is enlarged until a centroid is found, and the elbow position is decided and validated by anthropometric measurements and previous position. Similarly, the lower arm position is estimated by using the elbow position as the reference, see Fig.3 (right).

V. Experimental Results

We demonstrate our volume reconstruction and pose estimation algorithms on a video sequence of two-person posing

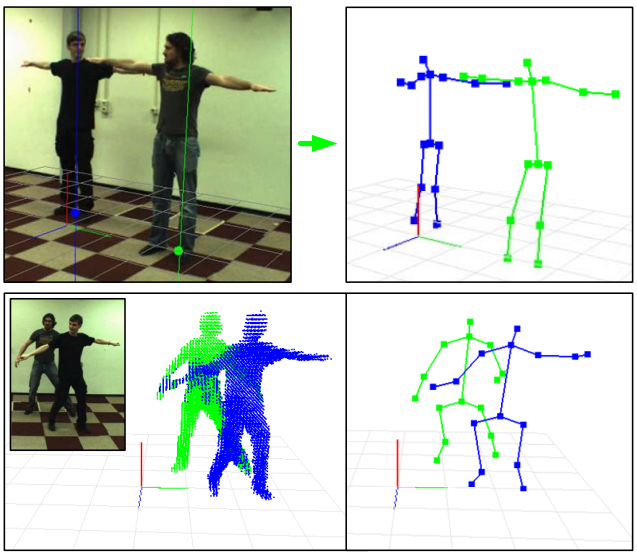

Figure 4. Example pose estimation for two persons.

captured by 4 calibrated Basler PiA cameras (644×484), placing in the frontal, left, right and side of the persons, see in Fig.1 (b). This sequence simulates the interactions of two persons with a computer, without body contact with each other. The two persons standing separately at the beginning to provide individual foreground blobs in all views, enabling our approach to automatically initialize the appearance models and approximate the 3D ground plane positions. After T-pose are shown as in Fig.4 (top row) the 3D body models are initialized by the first iteration of body model fitting. In the following time-frame, while the persons constantly move, walk across each other and perform body poses like stretching arms, the body model fittings are performed to estimate the poses. Fig.4 (bottom row) shows an example volume reconstruction and pose estimation.

With only fours views in our experimental setup, the pose estimation can be more difficult when the number of person increased (e.g. four or more). Then, inter-person occlusions become severe and in some cases, one or more persons cannot be viewed from any of the cameras. In this case, the quality of the volume will be improved using more cameras and properly placed around the scene. By using the previous body parts' location and the skeleton body model length and joints as the fitting constraints, the algorithm will be more robust.

VI. Conclusion

In this paper we proposed a novel framework for human pose estimation based on labeled voxels for multiple persons. Based on a novel technique for 3D tracking for multiple persons, we developed a voxel labeling technique based on measuring the distance to the reference position and considering inter-voxel connectivity inside the same volume. The quality of the estimated volumes is validated by successful indirect body model fittings, and our experiments show that the poses are properly estimated with high accuracy. Due to its efficiency, this framework is potentially suitable for various real-time multi-person HCI applications.

Acknowledgment

This research has been supported by the GATE (Game Research for Training and Entertainment) project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

References

[1] C. Tran and M. Trivedi, "Human body modelling and tracking using volumetric representation selected recent studies and possibilities for extensions," *ICDSC*, pp. 1–9, 2008.

[2] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and motion capture using voxel data," *Articulated Motion and Deformable Objects*, 2002.

[3] B. Michoud, E. Guillou, and S. Bouakaz, "Real-time and markerless 3d human motion capture using multiple views," *Human Motion - Understanding, Modeling, Capture and Animation*, 2007.

[4] B. Berendsen, X. Luo, W. Heurst, and R. C. Veltkamp, "Volumetric modeling of 3d human pose from multiple video," *SAMT Workshop on Semantic 3D Media*, 2008.

[5] Y. Iwashita, R. Kurazume, T. Hasegawa, and K. Hara, "Robust motion capture system against target occlusion using fast level set method," *ICRA*, pp. 168–174, 2006.

[6] L. Guan, J.-S. Franco, and M. Pollefeys, "Multi-object shape estimation and tracking from silhouette cues," *CVPR*, pp. 1–8, 2008.

[7] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *PAMI*, vol. 16, no. 2, pp. 150–162, 1994.

[8] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," *ECCV*, pp. 98–109, 2006.

[9] A. Komuravelli, A. Sinha, and A. Bishnu, "Connectivity preserving voxel transformation," *Discrete Applied Mathematics, Special issue for 12th International Workshop on Combinatorial Image Analysis, (in press)*, 2009.