Pose Estimation of Human Upper Body Using Multi-joint CG Model and Stereo Video Images

Kimio Hirao[*1], Atsushi Nakazawa[*1, *2], Kiyoshi Kiyokawa[*1, *2], and Haruo Takemura[*1, *2]

[*1]Graduate School of Information Science and Technology, Osaka University

[*2] Cybermedia Center, Osaka University

hirao@lab.ime.cmc.osaka-u.ac.jp, {nakazawa, kiyo, takemura}@ime.cmc.osaka-u.ac.jp

Abstract

We present a body pose estimation method for user interface and telecommunication applications, such as avatar control and desktop operations. We designed our method to meet the following conditions: marker-less, real-time and robust. The estimation is done with two-step matching. First, the acquired image is compared with the prepared CG model images. These model images are pre-generated by using both motion capture data and a camera parameter. When an input image is captured, image matching is done by evaluating the similarity between CG model images and the input images. The joint angles are then estimated by choosing the highest correlated image. Starting with an initial estimation result, joint angle parameters are refined through the iterative model-image synthesis and matching processes. Through this two-step pipeline, we can estimate fine joint angles very robustly in real time. We describe experimental results that show the validity of our approach. The matching process works at 10 frames per second.

Key words: Motion Capture, Matching, Refinement, Image Tracking

1. Introduction

Human body pose estimation is very necessary for computer graphics, user interfaces, security applications and so on. Several methods have been proposed for obtaining human body poses through both active and passive sensing methods [1]. 3-D model-based methods extract local image features and fit a given 3-D shape model to the features [2] [3]. 2-D appearance-based methods register the possible 2-D appearances of the target object and then find the best-matching one to the input image [4]. Shimada's method is estimating 3-D hand posture based on 2-D appearance [5]. Yoshimoto's method is using multi-view-based algorithms with multiple cameras [6]. D'Apuzzo's proposed a marker-less method using optical flow for upper body pose estimation with CCD cameras [7]. But this method is too slow to use for real-time applications. Zhu has proposed the method to estimate the 3-D head poses by using hybrid sensing through both depth and grayscale images [8]. This method however can only obtain the head poses. Our method is a passive vision method and it has the following characteristics: 1. marker-less, 2. real-time, and 3. robust. On the other hand, we don't need to estimate whole-body poses and decided to focus on only for upper body pose estimation. Using our method, in telecommunication application such as avatar chat system, each user can obtain the partner's pose with the partner's joint angles. Therefore communication cost will be slight.

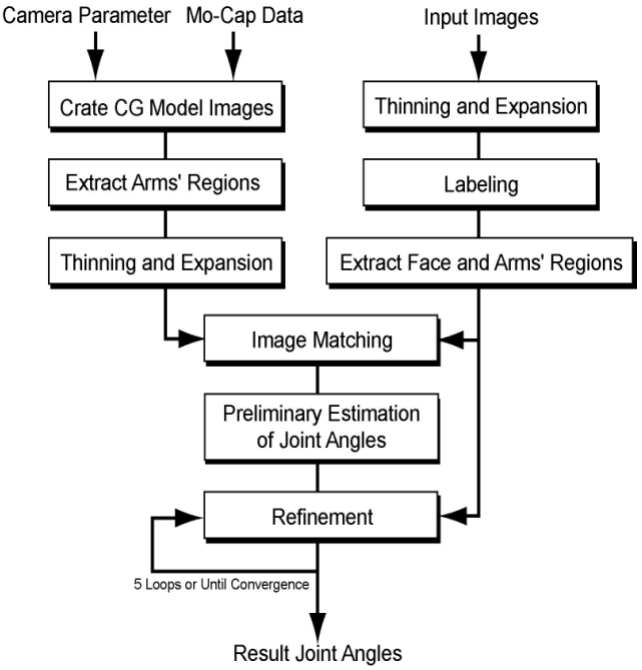


Fig. 1 Algorithm flowchart

2. Overview

The proposed algorithm is shown in Figure 1. Our method estimates both arms' poses using grayscale and depth images. The estimation is done with two-step matching.

First, an acquired grayscale image is compared with the

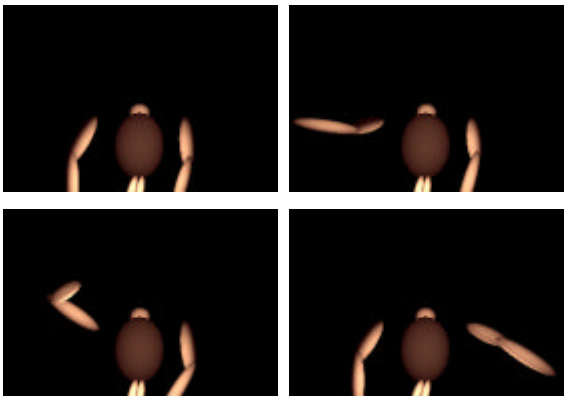


Fig.2 CG human model images

prepared CG model images. These model images are pre-generated by using both a camera parameter matrix and various joint angles acquired from motion capture data. Then image matching is done by evaluating the similarity between CG model images and the input image. Joint angles of both arms are initially estimated by choosing the most correlated CG model image. Starting with initial estimation result, joint angles are refined through the iterative model-image synthesis and matching processes. Through this two-step pipeline, we can estimate fine joint angles very robustly in real time.

2.1 Camera calibration

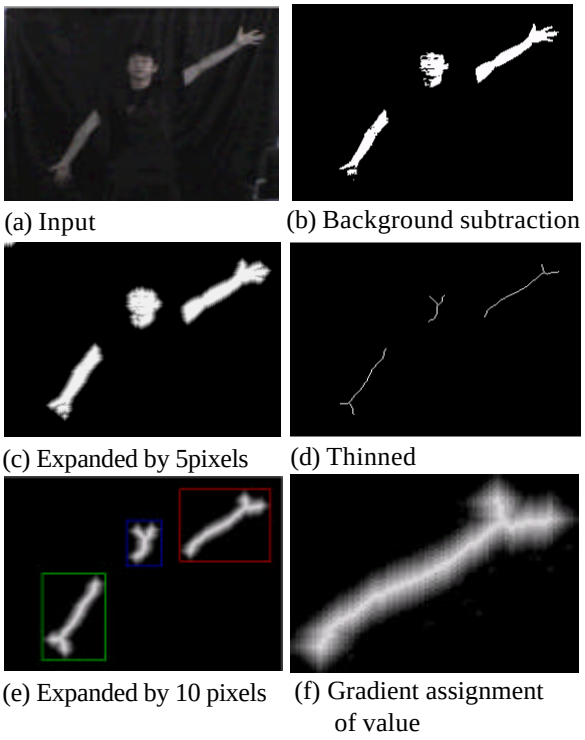
First, the camera parameter is obtained using a calibration box and image feature detection. The relatio- n between world coordinate system and camera coordin-ate system is shown by the following equation (1).

$$\begin{bmatrix} H_C X_C \\ H_C Y_C \\ H_C \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

The matrix C is called camera parameter, and this shows the relationship between camera coordinate (Xc, Yc) and world coordinate(X, Y, Z).

2.2 Generate model images

In advance, various poses of both arms are acquired as a set of pose data using optical motion capture (Oxford Metrics Vicon8) and the multi-joint CG model is generated. The sizes of body portions (links) are acquired from the subject. Here, the CG model has eight-degree-of-freedom: shoulders (3x2 DOF), and elbows (1x2 DOF). We set the waist position of the CG model as the origin of world coordinate system, and CG model images of various poses of both arms are generated (Fig. 2). Then each arm region is extracted from all CG model images. In this experiment, 4796 CG model images and 9592 each arm's images are generated.



(a) Input (b) Background subtraction

(c) Expanded by 5pixels (d) Thinned

(e) Expanded by 10 pixels (f) Gradient assignment of value

Fig. 3 Extraction from input

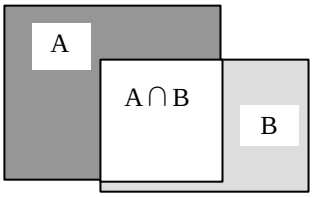


Fig. 4 Example of overlapping bounding boxes

2.3 Extract the body portions from input images

From captured images (Fig. 3a), a face's and both arms' regions are extracted by simple background subtraction and brightness threshold (Fig. 3b). Here, we suppose users are wearing black short-sleeved shirt, and the face's and arms' regions are brighter than background. Since their regions may be divided by background subtraction noises, each extraction region is expanded by 5 pixels (Fig. 3c), thinned (Fig. 3d), and expanded by 10 pixels again (Fig. 3e). Three rectangles shown in figure 3(e) show the bounding box of each human region. Next, the face's and arms' regions are extracted. The labeling is done for extraction regions and the COG (Center of Gravity) positions and the areas are calculated for each labeled regions. Considering these positions and areas, left and right arms and a face are determined.

2.4 Setting of values for body regions

For each arm in CG model images, the regions are similarly thinned and expanded by 10 pixels. In the last expanding process, the value 0 is set for background pixels, the highest value is set for the bone part of the

human region, and gradually lower values are set when it becomes close to the circumferential edge of background pixels. By using this gradient assignment of value, robust matching can be done. The image that expanded left arm's region of figure 3(e) is shown in figure 3(f). That region is displayed more brightly when its value is higher.

2.5 Matching

Image matching is done by evaluating the similarity between CG model images and the input image. This is calculated using the following equation:

$$eval = \frac{\sum_{x,y} \| p(x,y) - m(x,y) \|^2}{\sum_{x,y} (p(x,y)^2 + m(x,y)^2)} \quad (2)$$

where p(x, y): input image, m(x, y): CG model image. This equation means that the evaluation value decreases when the overlapping area between the input image and the CG model image is larger. Each evaluation value for all CG model images is obtained and the joint angles are estimated by choosing the CG model image that shows the smallest evaluation value.
To decrease the computational cost, the equation is applied to the particular CG model images satisfying the following conditions:

$$-\frac{S_{AandB}}{S_A + S_B - S_{AandB}} \quad (3)$$

$$\left| \frac{A.right - A.left}{A.top - A.bottom} - \frac{B.right - B.left}{B.top - B.bottom} \right| \quad (4)$$

$$\sqrt{d_x * d_x + d_y * d_y} \quad (5)$$

$$d_x = \frac{A.right + A.left}{2} - \frac{B.right + B.left}{2} \quad (6)$$

$$d_y = \frac{A.top + A.bottom}{2} - \frac{B.top + B.bottom}{2} \quad (7)$$

where A (B): a bounding box of the compared arm, S_A (S_B): the area of the bounding box, S_{AandB}: the overlapping area of bounding boxes A, B (Fig. 4). If each equation value is under appropriate threshold, the equation (2) is applied.

2.6 Refinement

Only the first image matching, joint angles that do not exist in motion capture data can't be estimated. So starting with this preliminary estimation result, joint angles are refined through the iterative model-image synthesis and matching processes. The outline of processing is shown in figure 5. Similar idea is seen in Moritani's method which intends the motion tracking of rigid object [9]. In our method, each joint parameter is slightly changed and differential images are generated.

But if we suppose all combinations of the joint parameter's movement, we have to evaluate too many differential images. So we suppose the movement of only one joint parameter for one refinement process. When we estimate a joint parameter, 8 differential images are generated in consideration of quaternion parameter combination. Those images are also compared with input image by the equation (2) and the joint angle is modified by choosing the most correlated differential image. Similarly, same processing is done for the shoulders' and elbows' joints. Totally, 32 differential images are generated to update all joint angles. This processing is repeated 5 times at a maximum for an input image until all joint angles converge.

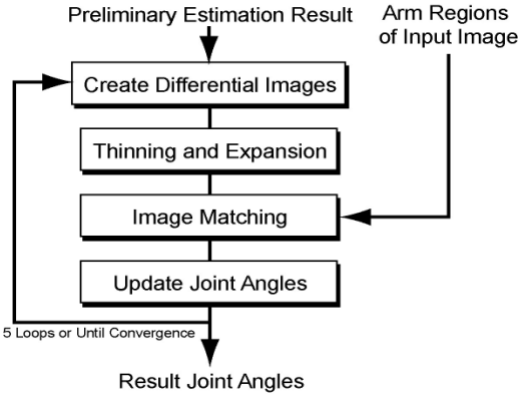


Fig. 5 Refinement flowchart

Fig. 6 Experimental environment

3. The Experimental result

The experimental environment is shown in Figure 6. Here, we use a PC (windows2000, CPU: 2.0GHz(Pentium4), RAM: 512MB) and Komatsu high speed stereo vision (FZ930). The experimental result is shown in Figure 7. Left images are input images and right images are output images. Without refinement process, our system works in 10 frames per second. With refinement process, it works in 2 frames per second. But refinement process gives more accurate estimation. Through the experiment, we can confirm this method works robustly in real time. But, we also found some failure cases such as figure 8, 9 where (a): input image, (b): extraction image, (c): output image. This method is weak for background subtraction noises and overlaps of the human regions.

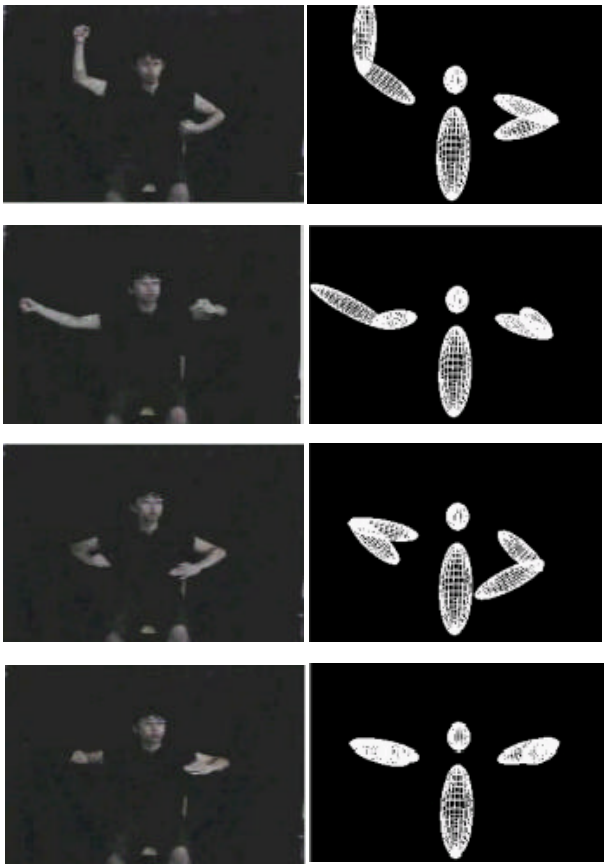
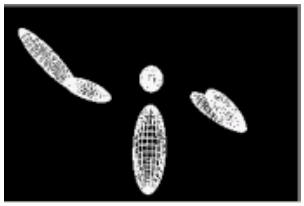


Fig.7 Experimental result

(a) Input image (b) Extraction image

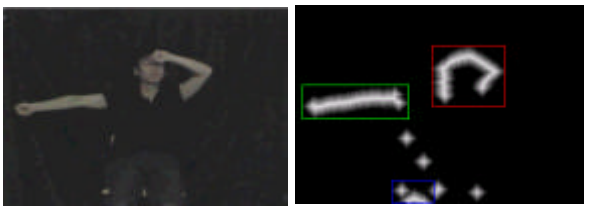


(c) Output image

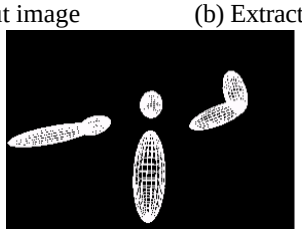
Fig. 8 Failure 1

4. Conclusion

In this paper, we have proposed the pose estimation method of human upper body by using the multi-joint model and stereo video image. The proposed method can robustly recover human upper body motion in real time by using motion capture based CG simulation. Using pre-generated CG model images, we can reduce the computational cost. At the moment, our implementation is not using depth images. As for future work, we would like to take into consideration the waist or the head.



(a) Input image (b) Extraction image



(c) Output image

Fig. 9 Failure 2

References

[1] Thomas B. Moeslund and Erik Granum, “A Survey of Computer Vis ion-Based Human Motion Capture,” Computer Vision and Image Understanding, Vol. 81, pp. 231-268, 2001.

[2] J. M. Rehg and T. Kanade. “Visual Tracking of High DOF Articulated Structures: an Application to Human Hand Tracking”. ECCV’94, pp. 35-46, 1994.

[3] D. Lowe. “Fitting Parameterized Three Dimensional Models to Images”. IEEE Trans., Pattern Anal.Machine Intell., vol. 13, No. 5, pp. 441-450, 1991.

[4] B. Moghaddam and A. Pentland. “Maximum Likelihood Detection of Faces and Hands”. Proc. of Int. Workshop on Automatic Face and Gesture Recognition, pp. 122-128, 1995.

[5] A. Imai, N. Shimada and Y. Shirai, "3-D Hand Pos-ture Recognition by Training Contour Variation", Proc. of 6th Int. Conf. on Automatic Face and Ge -sture Recognition pp. 895-900, 2004.

[6] Hiromasa Yoshimoto, Naoto Date, Daisaku Arita, Rinichiro Taniguchi, Satoshi Yonemoto: Vision-based Real-time Motion Capture System Using Multiple Cameras; IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems, Proc. MFI2003, pp.247-251, 2003.

[7] Nicola D’Apuzzo, “Human Body Motion Capture from Multi-Image Video Sequences,” Proc. of SPIE, Vol. 5013, Santa Clara, USA, pp. 54-61, 2003.

[8] Youding Zhu, K.Fujiwara, “3D Head Pose Estimat-ion with Optical Flow and Depth Constraints,” Proc. of 3DIIM, pp. 211, 2003.

[9] Takayuki Moritani, Shinsaku Hiura, and Seiji Inokuchi: Object Tracking by Comparing CG Images with Multiple Viewpoint Images; IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems, Proc. MFI2003, pp.241-246, 2003.