Three Dimensional Motion Estimation of a Human Body Using a Difference Image Sequence

KAMEDA Yoshinari, MINOH Michihiko, IKEDA Katsuo

Faculty of Engineering, Kyoto University
Yoshidahonmachi, Sakyoku, Kyoto, 606-01, Japan
kameda@kuis.kyoto-u.ac.jp

Abstract

We propose a model-matching method to estimate a motion of an human body from a difference image sequence. We assume that shape information of the human body and camera calibration parameters have been given to the system in advance. In the algorithm, parts which consists of the human body are classified into 3 modes: moving, stationary, and occlusion mode. When a part is in moving mode, the system matches the part's projection with a difference image. when it is in stationary mode, the system doesn't use the image. when in occlusion mode, the system predicts the motion from its locus and doesn't use the image.

1 Introduction

Motion estimation which aims to handle deformable objects is one of the important research in image understanding. Among the deformable objects, a human body has a significant value to be estimated because it is very useful to recognize a user's intension and instruction.

Recognition of motions of a human body is a challenging problem in computer vision. Rohr[4], Bharatkumar[3], Attwood[5] have contributed to solve this problem. Many researches use a stick or cylinder model and edges in images. However, human bodies are not stick in a strict sense. Moreover, edges in images usually contains not only useful ones but also many meaningless ones.

In this paper, we propose a human motion estimation method using a difference image sequence. A difference image is obtained from continuous two images which are taken by a fixed camera. Our method is based on a model matching method. A human model consists of 15 nodes each of which represents a part of a human body and has a joint attribute. In this paper, motion estimation is defined to estimate the joint angles in the model. In each frame of an image sequence, our model matching algorithm takes a node of the model one by one. We classify a status of the nodes into three modes; moving mode, stationary mode, and occlusion mode. A node is labeled with one of three modes by considering the relationship between the difference image and the projected region of the node on the image plane. We calculate joint angles according to the modes of the nodes at each frame. If a node is in the occlusion mode, the images cannot be used. To cope with such situation, we introduce inertia constraint such that the parts in a human body are generally rotated at a constant speed.

We assume that an image sequence does not include any moving object except for a human body and the camera calibration is done in advance and that the initial pose of the human body model is obtained from an other method such as [1][2].

2 Model

A human body is represented by an articulated object model. The model consists of 15 nodes. Each node corresponds to a part of human body. The nodes are connected to each other in a tree graph so that one node includes one joint that has at most 3 rotational axes. Joint rotational angles of the node i at time t are expressed by a vector $\mathbf{a}_i(t) = \{a_{ik}(t)\}$. The subscript k means the rotational axis of the joint rotational angle. The minimum and maximum value of $a_{ik}(t)$ are defined so that a human model prohibit unrealistic motion.

The motion is said to be estimated if the angles of all the joints are calculated for every frame in an image sequence.

3 Joint Rotation Considering Inertia

We assume that human motion generally undergoes inertia. In other words, we assume that a joint rotates with a constant speed if there is no factors to change the speed. In the real 3D world, the rotational speed of a joint may change to some extent. So, we use the formulation below to predict a joint rotational angle at time $t + \Delta t$.

$$|a_{ik}(t + \Delta t) - (a_{ik}(t) + v_{ik}(t)\Delta t)| \leq \Delta R \quad (1)$$

$v_{ik}(t)$ means the speed around the axis k of the node i at time t. The maximum rotational speed deviation ΔR can be determined depending on a kind of motion.

4 Node Mode and Difference Image

The points of our method are summarized into two issues; a node mode and segmentation of a difference image.

When the process estimates joint angles of a node the process decides which to use by its node mode, the difference image information or the joint angle prediction. A difference image is segmented into 4 types.

4.1 Node Mode

Suppose you observe a motion of a human body from the camera viewpoint. Human body parts can be classified into 3 modes: moving mode, stationary mode, and occlusion mode. To correspond these modes with the model, we classify the nodes in the model in the same way. The characteristics of these node modes are described below.

1. Moving Mode
   A part corresponding to the node labeled the motion mode can be seen from the camera. A difference image is available to estimate the joint angles of the node.
2. Stationary Mode
   A part corresponding to the node labeled the stationary mode can be seen from the camera. As the node is not moving, there is no area of it in a difference image.
3. Occlusion Mode
   A node in this mode can not be seen from the camera. In this case, we cannot get any information about the joint rotational angles from the images. So, we apply the inertia assumption to predict the joint angles.

Let $\mathcal{M}_t$ be a node set at time t where the nodes are in the motion mode, and $\mathcal{S}_t$ be a set of the stationary mode nodes and $\mathcal{O}_t$ be a set of the occlusion mode nodes.

4.2 Difference Image

Let $\mathbf{I}_{t_n}$ be an image of the frame n which is taken at time t_n and the images are taken with the time interval Δt. A pixel value s located at $\mathbf{x}$ at time t_n is defined as below.

$$s(t_n, \mathbf{x}) = |p(\mathbf{I}_{t_n}, \mathbf{x}) - p(\mathbf{I}_{t_{n-1}}, \mathbf{x})| \qquad (2)$$

$p(\mathbf{I}_t, \mathbf{x})$ means the pixel value at $\mathbf{x}$ in the image $\mathbf{I}_t$. A difference image is segmented into two regions according to $s(t_n, \mathbf{x})$.

- Stationary Region $\mathbf{S}_{t_n} = \{\mathbf{x} | s(t_n, \mathbf{x}) = 0\}$

  This region consists of the projections of the nodes in $\mathcal{S}_{t_n}$ and a background.

- Moving Region $\mathbf{M}_{t_n} = \{\mathbf{x} | s(t_n, \mathbf{x}) \neq 0\}$

  Moving region $\mathbf{M}_{t_n}$ consists of three regions; generated moving region $\mathbf{G}_{t_n}$, continuous moving region $\mathbf{K}_{t_n}$, and vanishing moving region $\mathbf{V}_{t_n}$. These three regions are defined below. In the formulations, $\mathbf{P}(\mathcal{A})$ means union of the projections of all the nodes in the node set $\mathcal{A}$ onto the image plane.

$$\begin{aligned} \mathbf{G}_{t_n} &= \mathbf{P}(\mathcal{S}_{t_{n-1}}) \cap \mathbf{P}(\mathcal{M}_{t_n}) && (3) \\ \mathbf{K}_{t_n} &= \mathbf{P}(\mathcal{M}_{t_{n-1}}) \cap \mathbf{P}(\mathcal{M}_{t_n}) && (4) \\ \mathbf{V}_{t_n} &= \mathbf{P}(\mathcal{M}_{t_{n-1}}) \cap \mathbf{P}(\mathcal{S}_{t_n}) && (5) \end{aligned}$$

5 Algorithm

This section describes the model matching algorithm we proposed. It utilizes the node modes and takes the relationship between the difference images and the projections of the human model into consideration. The process deals with a node one by one at each frame. The flowchart of the matching algorithm for a node is as shown in Figure 1.

Suppose the process comes to the frame n.

First, check whether the current node is in the occlusion mode or not. If it is in the occlusion mode, the joint angle is estimated according to the prediction using the inertia discussed before.

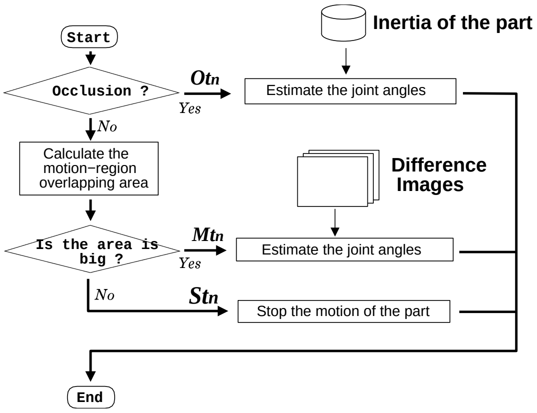

Start
Occlusion ?
Otn
Yes
Estimate the joint angles
Inertia of the part
No
Calculate the motion−region overlapping area
Difference Images
Is the area is big ?
Mtn
Yes
Estimate the joint angles
No
Stn
Stop the motion of the part
End

Figure 1: Flowchart of processing a node at each frame

Then, search $\hat{\mathbf{a}}_p(t_n)$ which maximizes the area size $F[\mathbf{a}_p(t_n)]$ by varying the joint angle $\mathbf{a}_p(t_n)$. If the difference image is the first one in the image sequence (that means $n = 2$), $F[\mathbf{a}_p(t_n)]$ is formulated as the equation (6). Otherwise, it is defined as the equation (7).

$$F\ [\mathbf{a}_p(t_2)] = f\left(\left(\mathbf{G}_{t_2} \oplus \mathbf{T}(\mathbf{a}_p(t_2))\right) \cap \overline{\left(\mathbf{H}(\{\mathbf{a}(t_1)\}) \cup \mathbf{C}_{t_2}\right)}\right) \qquad (6)$$

$$F\ [\mathbf{a}_p(t_n)] = f\left(\left(\mathbf{M}_{t_n} \oplus \mathbf{T}(\mathbf{a}_p(t_n))\right) \cap \overline{\mathbf{C}_{t_n}}\right) \qquad (7)$$

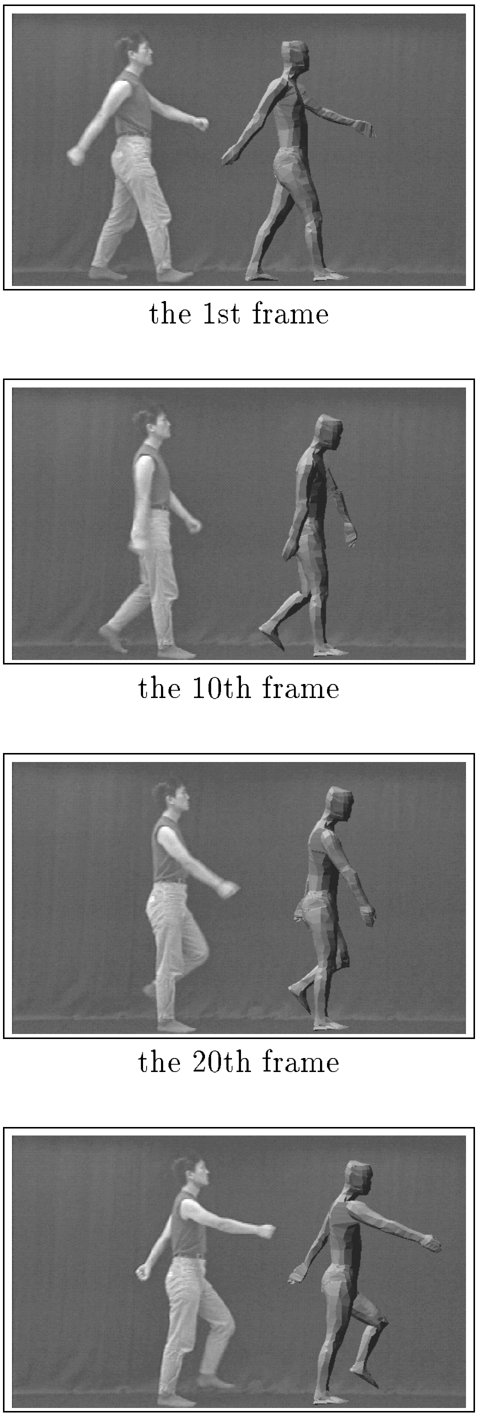
In the formulations, $\mathbf{T}(\mathbf{a}_p(t_n))$ indicates the region projected by the node p at time t_n. $\mathbf{C}_{t_n}$ is an union region of the projections of the nodes that have been estimated at time t_n. And $\mathbf{H}(\{\mathbf{a}(t_1)\})$ is a projection of the human model at time t_1 that is given to the system in advance. $\oplus$ is an exclusive-or operator between two binary regions.

If $F[\hat{\mathbf{a}}_p(t_n)]$ is larger than the threshold value μ, the node p is determined to be moving, otherwise stationary. The threshold μ is introduced to remove the influence of the noise occurred in the calculation of the difference. If the node is determined to be in the moving mode, the joint angles of the node is set to $\hat{\mathbf{a}}_p(t_n)$. If the node is determined to be in the stationary mode, the joint angles are set to the predicted ones discussed in Section 3 and its rotational speed is reset to zero so that it stands still.

After processing all the nodes in the model at one frame, the model matching algorithm goes to the next frame and continues the process until it comes to the end of the image sequence.

6 Experiment

We applied our method to an image sequence with a male walking. The image size 600 by 360 pixels and the sequence includes 40 frames. ΔR is set to 6 degree and μ is $2,500$ pixels. The estimated result is shown in Figure 2. The result motion is displayed 100 cm right to the actual result location for the readers. Here you can see that our method keeps tracking the parts in spite of occlusions. As our method uses a complete human model and can obtain the joint angles from the image sequence, we can show the estimation result from any viewpoint. See Figure 3.



the 1st frame

the 10th frame

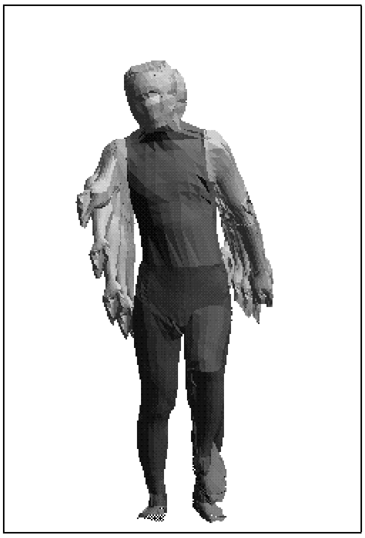
the 20th frame

the 30th frame

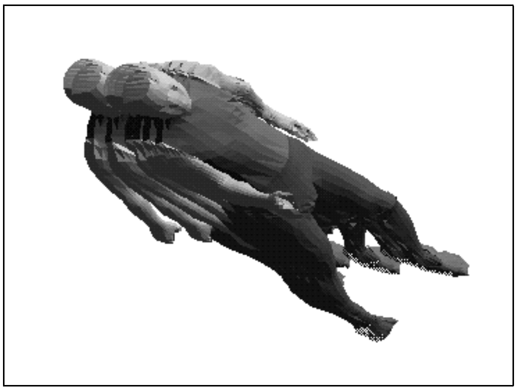
Figure 2: Result

7 Conclusion

In this paper, we propose a human motion estimation method using a difference image sequence. We classify a status of nodes in the model into three modes; moving mode, stationary mode, and occlusion mode. A node is labeled with one of three modes by considering the relationship between the difference image and the projected region of the part on



Front View



Slant View

Figure 3: Result from Another Viewpoint

the image plane. The process calculates joint angles by referring the modes of the nodes at each frame. Through the experiment, we showed that our method can keep tracking the parts of the human in spite of the occlusions.

References

[1] Y. Kameda, M. Minoh, and K. Ikeda. "Three dimensional pose estimation of an articulated object from its silhouette image," *Asian Conference on Computer Vision*, pp. 612–615, 1993.

[2] Y. Kameda, M. Minoh, and K. Ikeda, "A pose estimation method for an articulated object from its silhouette image (in Japanese)," *Trans. of the Institute of Electronics, Information and Communication Engineers*, Vol. D-II, 1995.

[3] A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, Qin Cai and J. K. Aggarwal, "Lower Limb Kinematics of Human Walking with the Medial Axis," *Proc. of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp.116-123, 1994.

[4] K. Rohr, "Incremental Recognition of Pedestrians from Image Sequences," *Proc. of the 1993 IEEE CVPR*, pp.8-13, 1993.

[5] C. I. Attwood and G. D. Sullivan and K.D. Baker, "Model-based Recognition of Human Posture Using Single Synthetic Images," Proceedings of the Fifth Alvey Vision Conference, pp.25-30 (1989).