A HUMAN MOTION ESTIMATION METHOD USING 3-SUCCESSIVE VIDEO FRAMES

KAMEDA,Yoshinari and MINOH,Michihiko

IMEL, Faculty of Engineering, Kyoto University,
Yoshidahonmachi, Sakyoku, Kyoto 606-01, Japan
Phone: +81-75-753-5995, Fax: +81-75-753-5965
E-mail: kameda@kuis.kyoto-u.ac.jp

Abstract

In this paper, we propose a human motion estimation method from three successive video frames. As it is no need to estimate a motion of a person when he stays still, we propose to use a *double-difference image* to get motion information from the video frames and estimate a pose partially on the region where motion feature is detected. A double-difference image is obtained by an AND operation between successive two difference images. We implemented a motion estimation system that executes model matching on a double-difference image.

1 Introduction

Automatic input and analysis of human motion have attracted attention from virtual reality and human interface researchers. The analysis by sensors or special devices were implemented and users find it useful these days. However, such special devices tend to impose a burden to measured persons and are not common in our daily life.

Computer Vision, especially using one camera, becomes main technical trend for human motion estimation and many researchers have contributed to it [5], [6], [7], [3], [4]. We have proposed to estimate human motion based on precise human shape model where input video frames are binarized in advance [1], [2].

On estimating a human motion by observing video frames, it is important what kind of features to be extracted from the video frames. We have concentrated on silhouette images in our previous research [2], but it is not easy to obtain the silhouette when the background or light conditions changes. In this paper, we propose to use difference images, because the difference operation derives motion information of the object in the images. If the human body does not move the difference value becomes zero, which means it is not necessary to estimate its motion.

Our method introduces a model matching method. A human body model consists of several solid objects each of which represents a part of a human body and has a joint attribute. Here, motion estimation is defined to estimate the value of the joint angles in the model.

We conducted the experiments on the real video frames. The results show that our method can keep tracking the motion of the human body.

2 Double-difference image

We propose to use a *double-difference image* to get motion regions from video frames and estimate a pose partially on the region where the motion regions is detected. The motion region is a region where a pixel value changes. A double-difference image is obtained by AND operation between successive two difference images.

It is assumed that the video frames do not include any moving object except for a human body.

We make a double-difference image from three successive frames in an video stream(Figure 1). First, we generate two difference images from corresponding two successive images ('t-1' and 't', 't' and 't+1'). Then we binarize the difference images and execute AND operation on these two images. We call a resultant binary image a *double-difference image*.

As a double-difference image is a product of two difference images, it tends to include isolated noise pixels. These pixels disturbs motion estimation described later in Section 4. Therefore, each 4 by 4 pixels in the double-difference image is grouped into one square block. A block is marked true if more than half of the pixels in the block is true. This process not only prevents noise but also reduces the computation cost in the image processing. Then the system removes isolated blocks to get rid of slight changes in the video frames. The motion regions consists of the pixels whose value is true in the remained blocks.

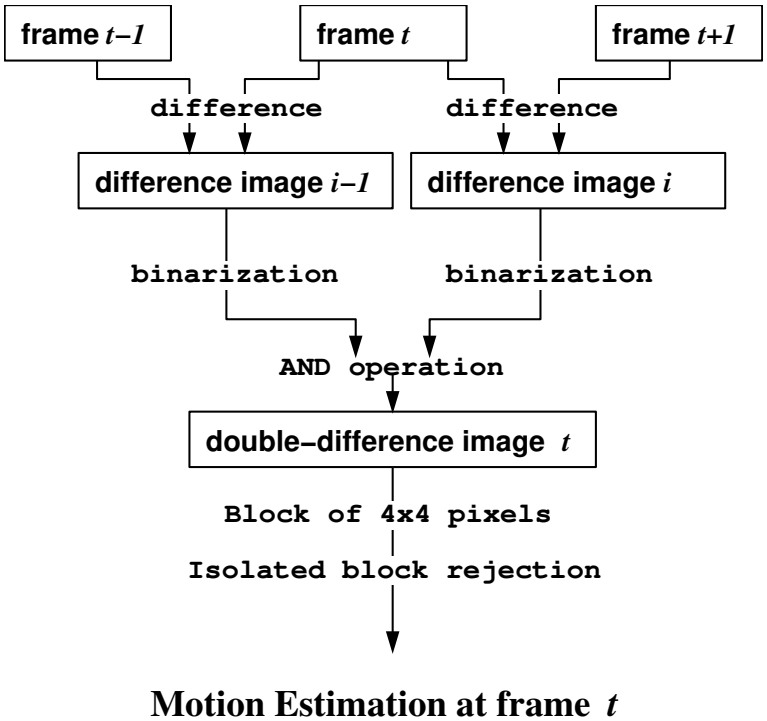


Figure 1: Double-difference Image Generation

A double-difference image has two good features. One is that motion regions on the double-difference image keeps the shape of the human body at time 't'. Regions on a normal difference

image do not express the shape of the object because it is a mixture of the object shape on the image plane at time 't-1' and that at time 't'. For example, consider a rectangle object transition in Figure 2. In the left, extracted shape is a combined contour (thick line) of that of time $t-1$ and time t. The right figure shows a double-difference image and the AND operation keeps the original shape at time t.

The other feature is that it is easy to detect whether the current frame contains motion information or not. If motion regions on a double-difference image are small or do not exist, it indicates that the human body stands still and it is no need to estimate the pose in that frame(Figure 3).

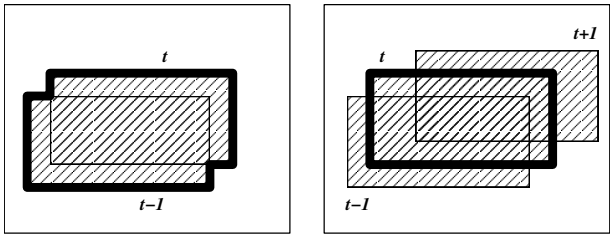


Figure 2: Extracted Shape on Difference Images

3 Human Model

A human body is represented by an articulated object model which consists of several solid body objects. A solid object corresponds to a part of the human body and has a joint to the adjacent body object. The body objects are connected to each other in a tree structure. According to the tree graph structure, pose of one body object is defined directly by the joint angles that is a attribute of the joint connecting it to a parent body object in the tree. A joint has at most 3 axes to rotate. Joint angles of the body object i at time t are expressed by a vector $\mathbf{a}_i(t) = \{a_{ik}(t)\}$. The subscript k means the axis of the joint rotation. The minimum and maximum joint angle are defined so that a human model does not come to an unrealistic pose. Figure 4 shows a human body model consisting of nine parts. The pelvis corresponds to the root body object. Black dots on the model represents the joints. For example, a joint of the head body part locates near the mouth in the figure. Dotted lines denotes the tree structure of this model.

In this way, the motion is defined by the time varied angles of all the joints for a certain period.

4 Motion Estimation Algorithm

We assume that the camera calibration is done in advance and that the initial pose of the human body model is obtained from the method we have proposed in previous paper [2].

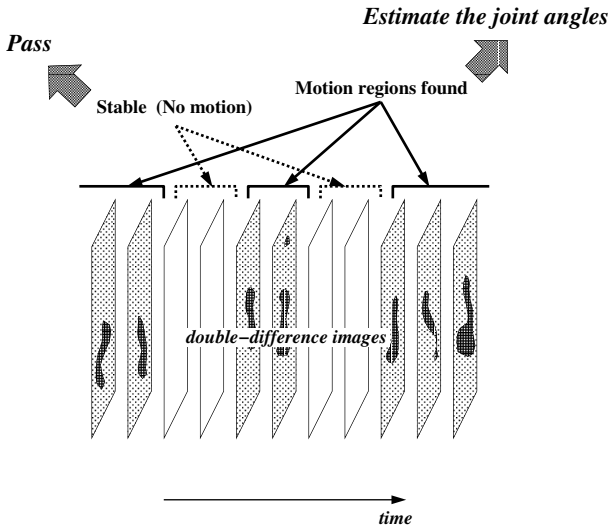


Figure 3: Frame Skip

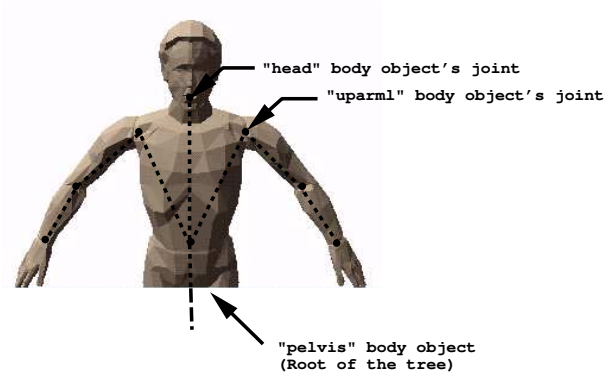


Figure 4: Human Body Model

Suppose now is time t. As shown in Figure 3, if the double-difference image at time t contains few motion regions (white frames in the figure), the joint values are all fixed as those in the previous frame.

On the other hand, if the double-difference image t contains motion regions(gray frames in the figure), it implies the human body has changed its pose. In this case, the model matching algorithm processes a body object in the model one by one, from the root position to the leafs of the tree structure in the model. We classify a status of the body objects into three modes; *moving mode*, *stationary mode*, and *occlusion mode*. Each body object is labeled with one of the three modes by observing the relationship between the motion regions and the projected region of the body object on the image plane. We calculate joint angles according to the modes of the body objects at each frame(Figure 5). If a body object is hidden by other body objects, it is determined to be in the occlusion mode and the image plane is not referred. To cope with such situation, we introduce inertia constraint such that the body objects in a human body are generally rotated at a constant speed[1].

5 Experiment

We implemented this method and applied it to real video frames. The human model used is shown in Figure 4. The video camera is set on the top of the desk bookshelf and focus on a person sitting in front of the desk. Camera position, camera direction, focus length are measured in advance. We constructed this system on SGI Indy (R4400 200MHz).

Figure 6 shows an example of the image processing. Figure (a) to (e) were taken at different moment. Figure(b) includes many noise pixels and they are rejected through (c), (d). In Figure (b), (c), (d), the pixels whose value is true in the double difference image express the original pixel value in the input frame. Final output Figure(e) passed to the motion estimation system includes small regions because each pixel in the motion regions is shown there.

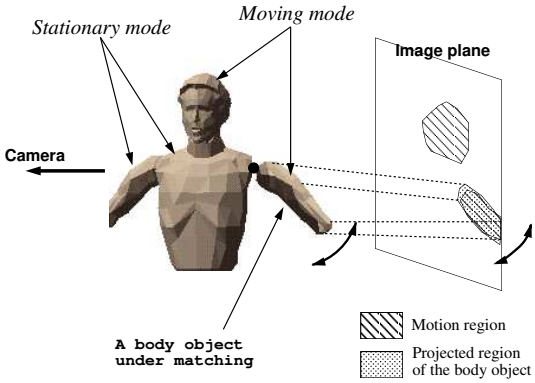
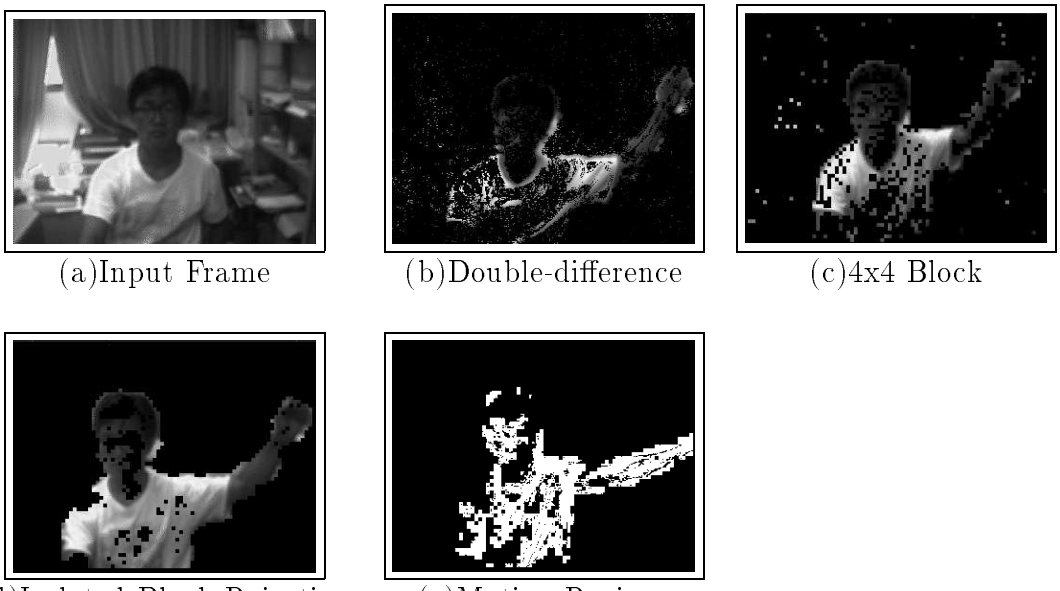


Figure 5: A Body Object under Matching Process



(a)Input Frame (b)Double-difference (c)4x4 Block

(d)Isolated Block Rejection (e)Motion Regions

Figure 6: Image Processing

The system generated 4.40 frames per second in average when the input video frame size was set to 320 pixels by 240 pixels and without motion estimation process. This frame rate depends on the number of motion regions in the video frame. See Table 1.

In the motion estimation experiments, we assumed that maximum rotation speed at each joint is 10 degrees per second. The experiments succeeded in simple exercises like waving hands. But there may exists difficulties in the present system. One of the reasons of the estimation failures is that one video frame consists of two fields. Since the current system merges two fields in one frame, shape of the motion regions do not reflect the correct shapes. Another reason of the estimation failures is that a body object once covers the incorrect motion regions, there is

no way to update the joint angles to the right values.

Through the experiment, the average rate of the stable double difference images for all the frames comes to 37.2%. This rate also depend on what the person is working in the office which cause the rate to be high.

The current system can process only 1.1 frames per second in average. Speed-up and computation reduction is our future work.

Table 1: Frame Rate for Image Process

	Average	Max	Min
frames/s	4.40	4.83	3.88

6 Conclusion

We have proposed a human motion estimation method from three successive video frames. As it is no need to estimate a motion of a person when he stays still, we propose to use a *double-difference image* to get motion information from video frames and estimate his motion only if the motion regions exists on the double-difference images. We implemented a motion estimation system that executes model matching on a double-difference image and showed its abilitiy.

References

[1] Y. Kameda, M. Minoh, and K. Ikeda, "Three Dimensional Motion Estimation of a Human Body Using a Difference Image Sequence," *Asian Conference on Computer Vision*, Vol. II, pp. 181-185, 1995.

[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.J-79-D-II, No.1, pp.26–35, 1996.

[3] James M. Rehg and Takeo, Kanade, "DigitEyes: Vision-Based Hand Tracking for Human-Computer Interaction," *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects*, pp.16-22, 1994.

[4] Francisco J. Perales and Juan Torres, "A System for Human Motion Matching between Synthetic and Real Images Based on a Biomechanic Graphical Model," *Proc. of the Workshop on Motion of Non-Rigid and Articulated Objects*, pp.83-88, 1994.

[5] 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.

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

[7] 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).