COMPUTER VISION BASED RECOGNITION OF INTERACTIONS BETWEEN HUMAN BODY AND OBJECT

Masumi Kobana and Jun Ohya
Graduate School of Global Information and Telecommunication Studies, Waseda University, Bldg. 29-7, 1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo 169-0051, Janan.
* *The first author is currently at NEC Corp. Japan.*

Abstract: This paper proposes a computer vision based method that recognizes interactions between human body and object. In two successive frames in a video sequence, our function based contour analysis method checks whether the silhouettes of the human body and object overlap. If they do not overlap, our method judges that the human body and object does not interact. If the two silhouettes overlap, our method checks whether the motion vectors obtained from the areas for the human body and object coincide. If they coincide, our method judges that the human body and object are interacting. Some experimental results show the effectiveness of the proposed method.

Key words: computer vision, recognition of interactions, silhouette analysis, motion vector

1. Introduction

Analyzing human motion is very important for a variety of applications. In industries that produce digital cinemas or video games, motion capture systems are frequently used for creating animations of human characters. In most of motion capture systems, sensing devices are attached to a human body so that the person's motion can be measured, but

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35660-0_65

R. Nakatsu et al. (eds.), *Entertainment Computing*

this contact type method for motion capture is cumbersome for humans and limits application areas.

Most of existing computer vision based approaches for analyzing human motion dealt with situations in which a person is independent of the environment that surrounds the person. Some of recent projects study how to track multiple persons in a scene, but not many works on recognizing interactions between a person and an object in a scene can be seen. Computer vision based method for recognizing interactions between a person and object can be applied to future type arcade games in which participants' physical movements as well as interactions with an object are used to control the games. I. Haritaoglu et al. [1] presented a method that identifies if a walking person is carrying an object, but the method assumes the person's periodical movements. This paper proposes an algorithm that does not need such an assumption and can judge whether the person interacts an object in the scene.

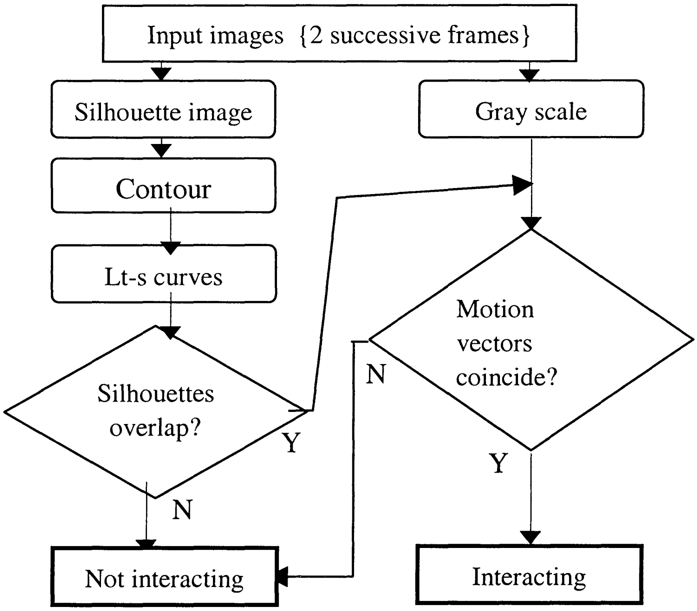


Fig1. Proposed algorithm

The authors have developed a method that estimates a person's postures by analysing the contour of the person's silhouette [2]. The proposed method improves the contour analysis so as to judge whether the person is interacting with an object. Section 2 elaborates on the proposed algorithm. The effectiveness of the proposed method is demonstrated by experimental results in Section 3.

2. Algorithm

2.1 Basic idea

Figure 1 shows a block diagram of the proposed algorithm. The basic strategy is as follows. In two successive frames, our function based method [2] analyzes the contours of the silhouettes to check whether the silhouettes of the human body and object overlap. If it turns out that the two silhouettes do not overlap, the system judges that the human body and object are separate. If it turns out that the two silhouettes overlap, it is necessary to check whether the human body and object are actually touching, because there might be cases in which the two silhouettes overlap without interactions between the human body and object. For this check, motion vectors obtained from the original gray-level images are exploited.

2.2 Function based contour analysis

Here, our function based contour analysis method, which is applied to silhouette images of a human body, is outlined. In this paper, we assume that a video camera observes a person so that a video sequence is acquired. At each frame (original gray-level image) of the video sequence, the silhouette image is obtained by subtracting the original image from the background image and thresholding the subtracted image. Let f_{ij} denote a silhouette image, where *i* and *j* are the vertical and horizontal coordinates of the image, respectively, where f_{ij}=1 corresponds to a pixel within the silhouette of the human body, and f_{ij}=0 corresponds to the background. To obtain the centroid G of the silhouette accurately regardless of different poses of arms and legs, a distance transformation is applied to f_{ij} (specifically, to the pixels whose values are 1), where the distance transformed image is represented by dij, whose pixel values indicate distance values. Then, another operation that suppresses the contribution of arms is applied to dij, where the obtained image is represented by g_{ij}. The orientation θ of the upper half of the human body is obtained as the inclination of the principal axis of inertia of the silhouette in g_{ij}, where PAU denotes the principal axis of the upper half of the body.

Figure 2 defines our function, called Lt-s function [2], which is used to analyze the contour of a silhouette. In Fig. 2, let A be a pixel in the contour. Suppose that {s} is a set of çontour pixels aligned counter clockwise. The Lt-s curve function is defined by the following equation:

$$Lt(s)=\sqrt{p^{2}{}_{t}(s)+g^{2}{}_{t}(s)} \qquad (1)$$

where $\boldsymbol{p_t (s) = PA}$ and $\boldsymbol{g_t (s) = GA}$.

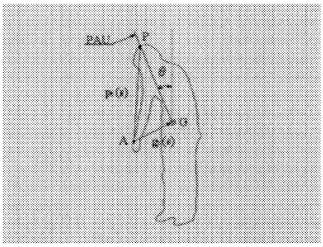
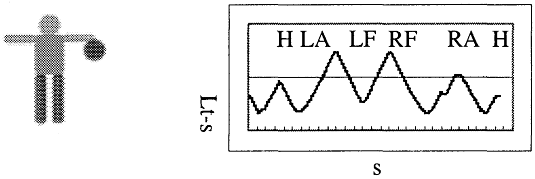


Fig. 2 Definition of Lt-s function



<Significant points> H:head, LA:left arm, LF:left foot, RF:right foot, RA:right arm

Fig. 3 Lt-s curves and significant points

Figure 3 shows an example of the Lt-s curve, which is obtained by plotting the value of the Lt-s function for each *s*. Local maxima in the curve correspond to significant points; in Fig. 3, H, LA, LF, RF, RA indicate the head top, the tip of the left hand, the tip of the left foot, the tip of the right foot, and the tip of the right hand, respectively. Note that if the person holds an object, the Lt-s curve changes, and the changes depend on the shapes of the objects. Basically, the proposed method utilizes this property of the Lt-s function.

2.3 Judging whether the silhouettes of the human body and object overlap

To judge whether the silhouettes of the person and object overlap, the Lt-s curves obtained from two successive frames are analyzed. The proposed method classifies the two successive frames into one of the following three cases: (1) from "non-overlap" (1st frame) to "overlap" (2nd frame), (2) from "overlap" to "non-overlap", and (3) no change (keep either "overlap" or "no-overlap" in the two frames). In the two frames, suppose that *a(s)* and *b(s)* are the Lt-s curves in the first and second frames, respectively. As shown in Fig. 4 (a) and (b), d is obtained as the difference in the number of contour pixels in *a(s)* and *b(s)*. Let N_o be the number of contour pixels of the object. Specific procedure for the above-mentioned classification is performed by evaluating the following conditions one after the other; that is, if a condition is satisfied, the judgement associated with that condition is made, otherwise the next condition is evaluated.

1 . If $|d-N_o| \leqq a$, then (1) (from "non-overlap" to "overlap").
2 . If $|d+N_o| \leqq a$, then (2) (from "overlap" to "non-overlap").
3. Otherwise$(a \leqq ||d|-N_o|)$, (3) (no change).

where a is a threshold (tolerance).

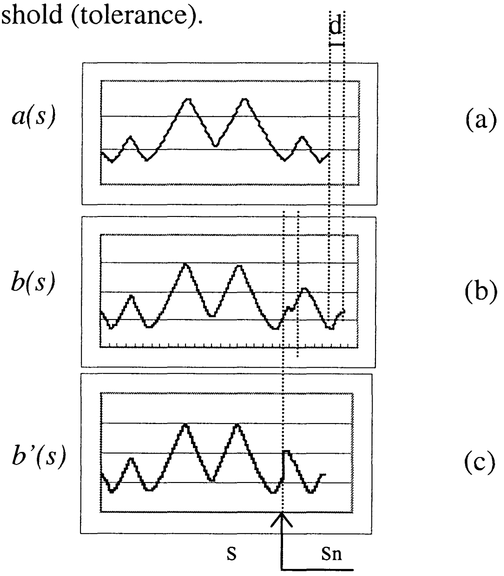


Fig. 4 Locating object range using cross-correlation

Next, it is necessary to locate the object in the contour of the silhouette if it turns out that the silhouettes of the human body and object overlap. The locating process is performed in case of the above-mentioned Condition 1 or Condition 3 (could be for multiple two successive frames) after Condition 1. For this, we utilize cross-correlation between *a(s)* and *b(s)*.

Note that *a(s)* and *b(s)* are normalized using their average values such that

$$a'(s) = a(s) - \frac{1}{N}\sum_{s} a(s) \qquad (2)$$

where N is the number of the contour pixels of *a(s)*. Similarly, *b'(s)* is calculated by normalizing *b(s)*. The cross-correlation value R between *a'(s)* and *b'(s)* is calculated by the following equation.

$$R(k) = \frac{1}{N-k} \sum_{n=0}^{N-1-K} a'(n)b'(n+k) \qquad (3)$$

where $k = 0, 1, ..., K$. In the proposed method, k is fixed to zero in Eq. (3) (thereby, $K=0$), because the point $s=0$ in *a(s)* and *b(s)* corresponds to the same point, namely the head of the top. If *a(s)* and *b(s)* match well, the value of *R* tends to become large.

Figure 4 explains the locating method using the cross-correlation in Eq.(3). As shown in Fig. 4 (a) and (b), *b(s)* is longer than *a(s)* by *d*; this means that the range that corresponds to the object in the contour has *d* pixels. Thus, *a(s)* and the Lt-s curve that is created by removing the range that corresponds to the object from *b(s)* should match well. The search process is performed as shown in Fig. 4(c); that is, the *d* pixels starting from the pixel s_n in {s} is removed from *b(s)* so that *b''(s)* is obtained after the normalization. By Eq.(3), the cross-correlation *R* between *a'(s)* and *b''(s)* is calculated. This calculation is done for all the pixels in {s} to locate the position (range) that gives the largest *R*.

2.4 Motion vector based judgement for interactions between the human body and object

The contour analysis described in Section 2.3 cannot judge whether the human body and object are actually touching, because two separate objects might be observed in the image as if the silhouettes of the two objects overlap. If the human body and object are touching, they could move together. In the proposed method, motion vectors obtained from the original gray-level images are utilized to judge whether the motion vectors of the human body and object coincide.

At each frame of two successive frames, gray-levels are preserved in the silhouette while the background area is assigned a uniform gray-level. Each of the two images obtained by this operation is divided into macro blocks. A search area is set around a macro block B in the first frame. Among the macro blocks in the search area in the second frame, the macro block B' that has the smallest gray-level correlation value is determined as the position of the block B in the second frame. The motion vector of B is obtained from the positions of B and B'.

The motion vectors obtained by the method described earlier could be contaminated by noise components. In the proposed method, the motions of the human body (normally, hand) and object are approximated by the following affine transformation based method.

Suppose that the centroid of the hand or object to be tracked is the origin of the 2D coordinate system. The relationship between the coordinates $\mathbf{x}_k$

of the center of a macro block and the coordinates $\mathbf{x}_k'$ of the position of the macro block at the next frame is represented using affine transformation by Eq. (4).

$$\mathbf{x}_k' = \mathbf{A}\mathbf{x}_k + \mathbf{d} \tag{4}$$

To obtain the motion vector that is not affected by noise components, the errors defined by Eq. (5) is minimized for all the macro blocks in the hand and object by the least square method so that the rotation matrix **A** and the translation vector **d** are calculated.

$$\varepsilon = \left(\mathbf{x}_k' - \left(\mathbf{A}\mathbf{x}_k + \mathbf{d}\right)\right)^2 \tag{5}$$

We define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}_1' & \cdots & \mathbf{x}_K' \end{bmatrix}^T$$

$$\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ \mathbf{x}_1 & \cdots & \mathbf{x}_K \end{bmatrix}^T$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{d} & \mathbf{A} \end{bmatrix}^T$$

If the determinant of $X^T X$ is not equal to zero,

$$\mathbf{B} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y} \tag{6}$$

holds. By Eq. (6), **d** and **A** are calculated.

If the motion vectors obtained for the hand and object coincide, the hand and object move together; in other words, the person can be judged to be interacting with the object using the hand.

3. Experimental Results and Discussion

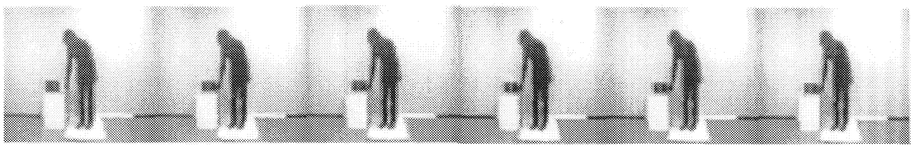
Experiments are conducted to confirm the effectiveness of the proposed method. For the experiments, video sequences that consist of 640 by 480 pixel frame are used. Telescopic search, which is characterized by its fast computation and is used for MPEG1's encoders, is utilzed for obtaining motion vectors from gray-level images. At each 640 by 480 pixel image, 16 by 16 pixel macro blocks are used, where the search area is set by extending each macro block area by 10 pixels vertically and horizontally.

At the first frame of the video sequence, the significant points such as the head top, hand tips and foot tips are located from the Lt-s curve. As shown in Fig. 5, at the second frame, the contour analysis method judged that the Condition 1 (from "non-overlap" to "overlap") is satisfied. By the cross-correlation based method, the object range, which is indicated by the white line in Fig.5, is located. From the third frame to sixth frame, the contour analysis method judged that the Condition 3 is satisfied (keep "overlap").

Then, we checked whether the hand and object are touching. From the first frame to the second frame, the motion vector of the object is almost zero, while the hand is moving. From the second frame to the sixth frame, the motion vectors of the hand and object are almost same; therefore, the hand

and object are touching in these frames. This result is same as the actual interaction.

At present, we can deal with only translational motions such as the one shown in Fig. 5. Remaining issues include that the method should be able to treat 3D motions including rotations. Variety of cases such as different postures and different objects should be dealt with.



Frame 1 2 3 4 5 6

Fig. 5 Video sequence used for the experiments for judging interactions between the hand and object

4. Conclusion and Future work

This paper proposes a method for judging whether the human body and object are interacting. Main results are as follows.

1. Contour analysis based on the Lt-s curve and cross-correlation based matching make it possible to locate the object range in the contour when the silhouettes of the human body and object overlap.
2. Motion vector detection using the telescopic search and affine approximation for object motions are useful for judging whether the human hand and object are actually touching.
3. Experimental results show a promising first step towards applying this method to future type game systems.

The proposed method should be improved so that persons' different postures and different objects can be dealt with.

Part of this work was supported by the Grant-in-Aid for Scientific Research (C) (2) 13650432, Japan Society for the Promotion of Science.

References

[1] I. Haritaoglu, et al., "Detection of People Carrying Objects using Silhouettes", International Conference on Computer Vision, p102-107, 1999.

[2] K. Takahashi et al., "Remarks on a real-time, non-contact, non-wear, 3D human body posture estimation method", Transactions of IEICE, Vol. J83-D-II, No.5, pp. pp.1305-1314• 2000.5. (in Japanese).

[3] ISO/IEC 11172-2 International Standard MPEG-1 Video 1993