A New Voting Algorithm for Tracking Human Grasping Gestures

Pablo Negri, Xavier Clady, and Maurice Milgram

LISIF - PARC, UMPC (Paris 6),
3 Galilee 94200 Ivry-sur-Seine, France
pablo.negri@lisif.jussieu.fr
{clady, maum}@ccr.jussieu.fr

Abstract. This article deals with a monocular vision system for grasping gesture acquisition. This system could be used for medical diagnostic, robot or game control. We describe a new algorithm, the Chinese Transform, for the segmentation and localization of the fingers. This approach is inspired in the Hough Transform utilizing the position and the orientation of the gradient from the image edge's pixels. Kalman filters are used for gesture tracking. We presents some results obtained from images sequence recording a grasping gesture. These results are in accordance with medical experiments.

1 Introduction

In gesture taxonomy [1], manipulative gestures are defined as ones that act on objects inside an environment. They are the subject of many studies in the cognitive and medical communities. The works of Jeannerod [2,3] are the references in this domain. These studies are frequently carried out to determine the influences of psychomotor diseases (Parkinson [4], cerebral lesions [5], etc.) on the coordination of grasping gestures. A typical experiment involves numerous objects, generally cylindrical, of varying size and position placed on a table. Subjects grasp the objects following a defined protocol. Active infrared makers are placed on the thumb, the index finger and over the palm. Vision systems, such as Optotrack, track the markers to record the trajectory performed during the test, in order to measure the subject responses to the stimulus.

In this paper, we propose a low-cost and less restrictive vision system, requiring only one camera and a computer. Many applications could be envisioned with this device, for example medical assistance [4,5], as natural human-computer interface (see [6] for more information on natural HCI) to control arm robots for grasping tasks [7], or virtual games.

We follow the experimental protocol described in [2,3,5,4]. The system is composed of a layout according to the specification of the Evolution Platform (EP), made of eight colored circles placed in a known geometry (see fig. 1). The hand in a grasping configuration (the index opposite to the thumb) moves in a horizontal plane with constant height (Z-axis). The camera is static and

J. Blanc-Talon et al. (Eds.): ACIVS 2005, LNCS 3708, pp. 130–137, 2005.

placed sufficiently far away in order to capture a complete view of the EP; with this configuration, we can consider the differences in Z-axis between hand points negligible. The 3D position of the finger are calculated with these assumptions and an iterative estimation of the camera pose [8] (see fig. 2).

Fig. 1. Camera view of the Evolution Platform

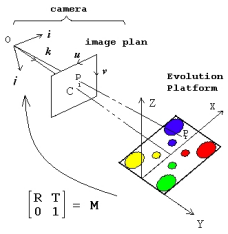

Fig. 2. Systems of reference for the camera and platform

In the next section, we describe the procedures to extract the finger positions in the image. Once a subregion of the image has been determined with a background subtraction algorithm, a skin color distance image is calculated for the extraction of the oriented edges of the hand. These edges were used in an original algorithm, the Chinese Transform (CT), for the segmentation and localization of the fingers. This approach is inspired from the Hough Transform. This algorithm allows the extraction of finger segments. Section 3 concerns the gesture tracking. We use a Kalman filter adapted to segment tracking. A simplified hand model makes it possible to determine the hand parameters generally used to analyze the grasping gesture. Section 4 presents some results obtained with our system. We conclude this article with some perspectives.

2 Detection of Hand and Fingers

2.1 Image Preprocessing

During the initial phase of the image preprocessing we used background subtraction to localize a sub-region in the image where the hand could be. We applied Stauffer and Grimson's Gaussian Mixture Model (GMM) method [9]. In our approach, we use the chrominance of the color space YC_bC_r instead of the RGB color space used by Stauffer and Grimson [9]. This permit us to ignore shadows of the hand. From this background subtraction, we obtain a binary image of the foreground pixels representing moving objects. We defined a search window including all the foreground pixels, which offers the advantage of reduced computing time for the following operations.

Secondly, the search window in the color image is converted to a skin color distance image. We emphasize the pixels with skin color chrominances, giving the maximum values in the skin color distance image. The converted RGB image, I,

into the YC_bC_r color space is referred to as I_{ybr}. Next, we calculate the inverse distance to the skin color, subtracting the chrominance channels b and r from I_{ybr} with the experimental values b_{skin} and r_{skin} respectively.

$$\overline{I_b} = |I_{ybr}(b) - b_{skin}|$$
$$\overline{I_r} = |I_{ybr}(r) - r_{skin}|$$
$$\overline{I_{br}} = \sqrt{\overline{I_b} + \overline{I_r}}$$

Finally, we obtained a skin color distance image (see fig. 5), defined by :

$$I_{sk} = 1 - \frac{\overline{I_{br}}}{max(\overline{I_{br}})}$$

2.2 The Chinese Transformation

The Chinese Transformation (CT) takes its name from Zhongguo, usually translated as *Middle Kingdom*, the mandarin name for China. The CT is a voting method: two points having opposite gradient directions vote for their mid-point. This method has the same basic principle as Reisfeld [10].

In the example of figure 3, two edge points e_1 and e_2 obtained from an image of an ellipse I_e, were defined with two parameters: the normal vector, n_i, and the position in the image, $p_i(x, y)$, with $i = \{1, 2\}$. Each normal vector was represented by the orientation of the gradient at this point. p_{12} was the segment drawn between the two points with p_v its mid-point.

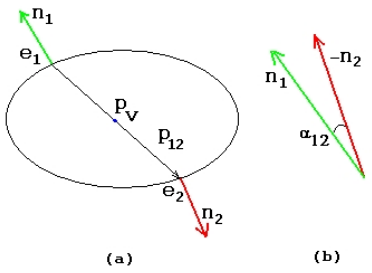

Fig. 3. (a) shows two edge points of the image with their normal vectors, (b) shows the two normal vectors superposed forming an angle α

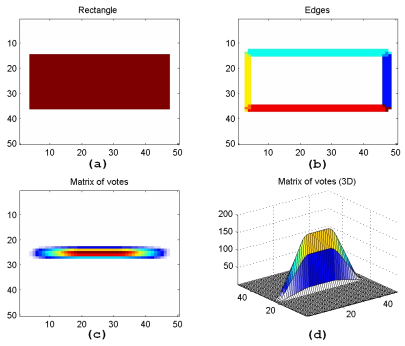

Fig. 4. Example of the CT for a rectangle. (a) original image, (b) oriented edges, (c) and (d) votes array in 2D and 3D spaces.

Superposing e_1 over e_2, we can compare their orientations. We say that e_1 and e_2 have opposite orientations if the angle α_{12} formed between n_1 and $-n_2$ satisfies the condition:

$$\alpha_{12} < \alpha_{threshold} \tag{1}$$

Then, the CT votes for p_v, the mid-point point of p_{12} if:

$$|p_{12}| < d \tag{2}$$

We create and increment an accumulator (votes array) with all the couples satisfying the conditions (1) and (2).

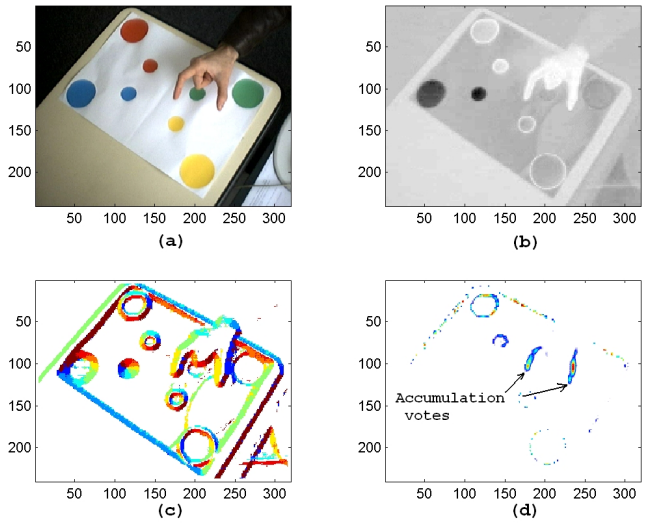

Fig. 5. Illustrations of Chinese Transform. (a) Original image, (b) Skin color distance image, (c) Oriented gradient image: each orientation is represented by a different color in the image, (d) Votes array.

The Fig. 4 is an example of the CT algorithm applied to a rectangle. The Fig. 4.a represents the original image. The Fig. 4.b shows the contour of 4.a and we can see the different orientations of the gradient in different colors. In practice, we sample the gradient orientations in $N = 8$ directions; this operation fixes a practical value for $\alpha_{threshold}$. The choix of this value depends on the application. With lower values there will be more pairs of voting points, increasing computing time. With higher values, we obtain much less voting points, losing information. The votes array (see Fig. 4.c and 4.d) is the result of applying the CT for all the edge points of Fig. 4.a with $d = 35$.

In our application, we take advantage of the form of the index finger and thumb (the two fingers forming the grip). Their parallel edges satisfy the distance and gradient direction conditions. The accumulation zones founded in the votes array can define the fingers regions (Fig. 5.d). A special function implementing a few variants of Hough Transform for segments detection is applied to the votes array (fig. 6). Each region's point votes, in the Hough space, with a value

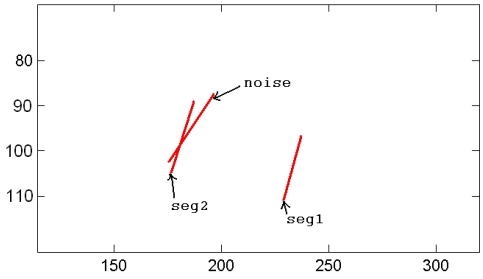

Fig. 6. The segments resulting from the CT

proportional to the quantity of votes in the accumulator. The resulting segments represent the finger regions.

Results of the CT can be compared with the morphological skeleton algorithm. But this method employs regions and has to deal with their usual defaults: holes presence, contour's gaps, partial occlusion, etc. The CT works around these problems with a statistical voting technique. In addition, it should be noted that CT could be useful in other contexts like detection of axial symmetries or the eyes localization in face images [11].

3 Gesture Tracking and Representation

Kalman filters [12] adapted to the segments makes it possible to both track the finger segments and eliminate the false alarms.

3.1 Segments Tracking

The objective is the tracking of segments belonging to the fingers in a sequence of images. This segments were obtained from the votes array of the Chinese Transformation and the application of the Hough Transformation. The parameters identifying each segment are (see fig. 7): $P_m(x_m, y_m)$, mid-point coordinates, l and θ, respectively the length and the angle of the segment.

Our system is made of three independent Kalman filters. Two scalar filters for the length and the orientation, and one vectorial filter for the position. If we consider constant speed, the state vectors are:

$$X^{P_m} = \begin{pmatrix} x_m \\ \dot{x_m} \\ y_m \\ \dot{y_m} \end{pmatrix} \qquad X^{l} = \begin{pmatrix} l \\ \dot{l} \end{pmatrix} \qquad X^{\theta} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}$$

We track all the segments in the image sequence. There are false alarms, but they disappear in the successive images. We keep and track the segments corresponding to the fingers. These segments follow the Kalman conditions and the constraints on the grip model.

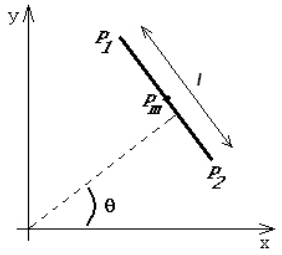

Fig. 7. Segment model

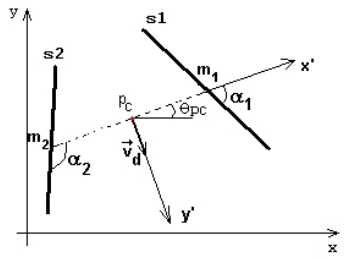

Fig. 8. Grip model

3.2 Grip Tracking

From two fingers segments, we model a grip (see Fig. 8) with the following parameters: p_c, mid-point of the segment m_1m_2; Θ_{pc}, angle that defines the inclination of the grip; v_d, unit vector which defines the grip's orientation and l_{12}, length of the segment m_1m_2.

We add two other parameters: the orientations α_1 and α_2 of the segments s_1 and s_2, calculated after a change of coordinate axis, from the (x, y) (see fig. 8.a) axis into the (p_c, x',y') axis related to the grip (see fig. 8.b).

The new state vectors for the grip tracking are: X^{p_c}, $X^{\Theta_{pc}}$, X^{α_1}, X^{α_2} and $X^{l_{12}}$. This representation was inspired from the studies on grasping gestures from [2,3,5,4]. We can easily observe the principal variables used in these studies: inter-distance between fingers (grip aperture), positions and orientations of the hand related to the scene (and to the objects), etc. In addition, we can define an articulatory model of the grip. By considering its geometrical model inversion, we can simulate a robotized representation of the grip in an OpenGL virtual environment.

4 Results

We apply the CT in a video sequence, taked from a webcam at 14 frames/second, composed of three stages. The first stage shows a hand going to grasp an imaginary object in the corner of the EP (see Fig. 9). The next stage shows the subject's hand puting the imaginary object down in the opposite corner of the EP. The final stage is the hand returning to the initial position. Here, we present only the result obtained for the stage 1.

Results of the CT on the first stage are shown in Fig. 10. In Fig. 10.a we can see all the segments recorded along the first 38 frames from the sequence. In the Fig. 10.b, the trajectory of the p_c point is shown. The grip aperture and hand velocity curves for the stage 1 are showed in Fig. 10.c and Fig. 10.d, respectively. According to Jeannerod, the arm mouvement to the target-object location of a normal subject is divided into 2 phases: a high-speed phase corresponding to the 75% of the movement of approach towards the object and a final low-speed phase [3]. In fig. 10.d we notice the first phase until the frame 15 characterized by a great acceleration and an increase in the distance inter fingers (fig. 10.c). Then,

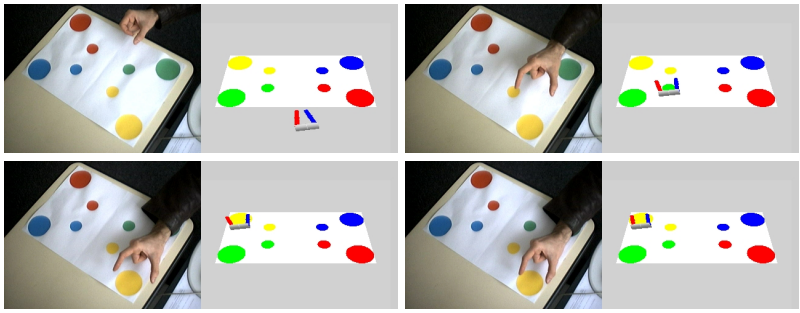

Fig. 9. This figure shows a picture sequence of the stage 1. On the left picture, we can see the camera view and, on the right image, the OpenGL environment with a virtual grip reproducing simultaneously the gesture.

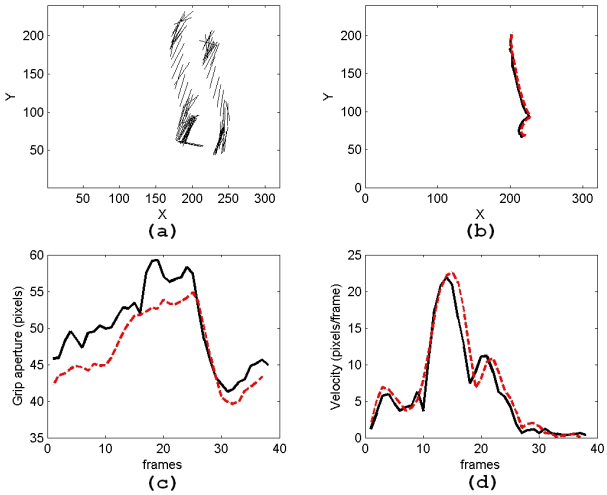

Fig. 10. Results for the stage 1: (a) all the segments. (b) trajectory of the point p_c. (c) distance between fingers (d) velocity of the hand. This figure show in dot points the hand labeled data for the curves (b), (c) and (d).

there are a deceleration of the hand while approaching to the object, reaching the final hand grip aperture.

This information can be used by the specialists in order to measure the patient's ability. They can also be useful for medical diagnostics.

5 Conclusion and Perspectives

Our goal was the tracking of the grasping gestures in a video sequence to detect psychomotor diseases. This article presented a system for the acquisition

and analysis of the human grasping gestures. It used a new method for fingers detection and localization, called Chinese Transform. This technique is a voting method inspired by Hough Transform. Kalman filters were used for gesture tracking. The obtained results were in accordance with observations of medical studies [2,3].

The next steps of our works will be oriented in the determination and the prediction of the grip points on an object. For that, we will have to analyze the gesture related to the intrinsic (form, size, etc) and extrinsic (position, orientation) object characteristics.

References

1. Pavlovic, V., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. PAMI **19** (1997) 677–695
2. Jeannerod, M.: Intersegmental coordination during reaching at natural visual objects. Attention and performance (Long J, Baddeley A, eds) (1981) 153–168
3. Jeannerod, M.: The timing of natural prehension movements. Journal of Motor Behavior (1984) 16:235–254
4. Castiello, U., Bennet, K., Bonfiglioli, C., Lim, S., Peppard, R.: The reach-to-grap movement in parkinson's disease: response to a simultaneous perturbation of object position and object size. Computer Exp. Brain Res (1999) 453–462
5. Hermdrfer, J., Ulrich, S., Marquardt, C., Goldenberg, G., Mai, N.: Prehension with the ipsilesional hand after unilateral brain damage. Cortex (1999) 35:139–161
6. Turk, M., Kolsch, M.: Perceptual Interfaces. In: Emerging Topics in Computer Vision. Prentice Hall PTR (2005)
7. Triesh, J., von der Malsburg, C.: Classification of hand postures agains complex backgrounds using elastic graph matching. Image and Vision Computing **20** (2002) 937–943
8. Oberkampf, D., DeMenthon, D., Davis, L.: Iterative pose estimation using coplanar feature points. Computer Vision and Image Understanding **63** (1996) 495–511
9. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR. Volume 2., Fort Collins, Colorado (1999) 22–46
10. Reisfeld, D.: Generalized Symmetry Transforms: Attentional Mechanisms and Face Recognition. PhD thesis, Tel Aviv University (1994)
11. Milgram, M., Prevost, L., Belaroussi, R.: Multi-stage combination of geometric and colorimetric detectors for eyes localization. In: ICIAP, Cagliari, Italie (2005)
12. Kalman, R.: A new approach to linear filtering and prediction problems. Transactions of the ASME - Journal of Basic Engineering (1960) 35–45