Hand Gesture Recognition in Camera-Projector System†

Attila Licsár[1], Tamás Szirányi[1,2]

[1]University of Veszprém, Department of Image Processing and Neurocomputing,
H-8200 Veszprém, Egyetem u. 10. Hungary
licsara@almos.vein.hu
[2]Analogical & Neural Computing Laboratory, Computer & Automation Research Institute,
Hungarian Academy of Sciences,
H-1111 Budapest, Kende u. 13-17, Hungary
sziranyi@sztaki.hu

Abstract. Our paper proposes a vision-based hand gesture recognition system. It is implemented in a camera-projector system to achieve an augmented reality tool. In this configuration the main problem is that the hand surface reflects the projected background, thus we apply a robust hand segmentation method. Hand localizing is based on a background subtraction method, which adapts to the changes of the projected background. Hand poses are described by a method based on modified Fourier descriptors, which involves distance metric for the nearest neighbor classification. The proposed classification method is compared to other feature extraction methods. We also conducted tests on several users. Finally, the recognition efficiency is improved by the recognition probabilities of the consecutive detected gestures by maximum likelihood approach.

1 Introduction

Video projection is widely used for multimedia presentations. In such situations users usually interact with the computer by standard devices (keyboard, mouse). This kind of communication restricts the naturalness of the interaction because the control of the presentation keeps the user in the proximity of the computer. In this paper we demonstrate an effective human-computer interface for a virtual mouse system in a projector-camera configuration. It would be more comfortable and effective if the user could point directly to the display device without any hardware equipment. Our proposed method interacts with the projected presentations or applications by hand gestures in a projector-camera system. For this purpose we use the image acquired by a camera observing the gestures of the speaker in front of the projected image. The system applies a boundary-based method to recognize poses of static hand gestures. The virtual mouse-based application is controlled by the detected hand poses and the palm positions. The virtual user-interface can be displayed onto the projected

† This paper is based on research supported by OTKA-T037829 of the Ministry of Education, Hungary.

background image, so the user controls and interacts directly with the projected interface realizing an augmented reality.

In the following we present related systems and give an overview of our work. In the next sections an overview of camera-projector systems and hand segmentation methods are described. Section 4 contains the gesture classification by several feature extraction methods. Finally, we describe an estimation method to increase recognition efficiency by collecting gesture probabilities in time.

2 Related Works

The aim of camera and projector based configurations is that the user interaction should be performed with the projected image instead of applying computer interfaces indirectly. A projector-camera pair is used to display the user interface on the projected surface (Fig. 1) where the camera acquires (camera image) the projected information (projected image) and the gestures of the user provide feedback about the interaction.

Usual methods apply standard white-boards or screens to display the information to the audience. The interaction can be induced by special hardware or vision-based methods. In SmartBoard [1] there are special display hardware devices with sensors e.g. to detect physical contact with the display or use laser pointer for the interaction with user interface by a vision-based system. BrightBoard [2] system uses a video camera and audio feedback to control the computer through painting simple marks onto the board. Other methods, like DigitalDesk [3], FreeHandPresent [4] apply hand gestures e.g. to navigate in a projected presentation by a restricted gesture set by counting and tracking fingers against to the cluttered background. The changing background disturbs the finger finding process so it defines a control area on a white background next to or above the projected surface. The projected image involves restricted background containing only figures and texts. This method applies finger resting on an item for 0.5 seconds as a “pick up” gesture. Magicboard system [5] applies camera and projector pair to get spatio-temporal activities on a white board by finger tracking method. In [6] system works with back-projection screen and can be used for mouse-cursor positioning by pointing with the arm. It determines the arm direction by 3D stereo computation and command is generated by voice signal.

Our goal is the development of a more natural interface in a camera-projector system using hand gesture analysis. The proposed system detects hand poses by shape analysis resulting in a larger vocabulary set in the communication. The correct shape detection is solved in the presence of a cluttered background.

3 System Overview

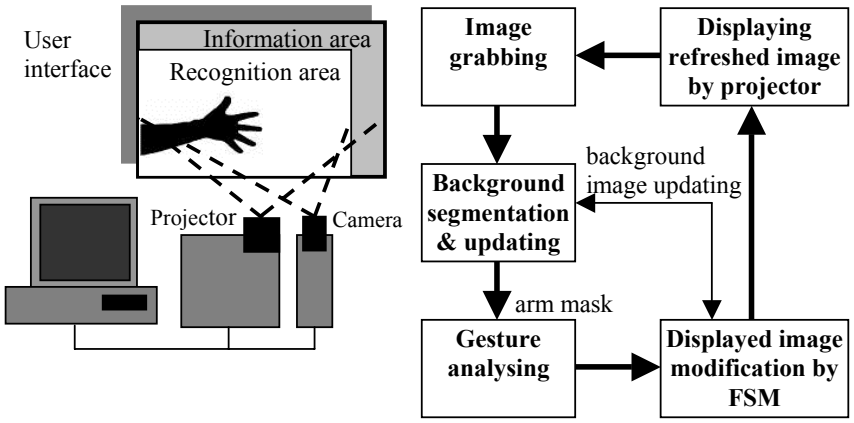

Fig. 1. System configuration and processing scheme

In our proposed method the arm and the forearm are segmented from the projected background. Our method uses a large gesture vocabulary with 9 hand poses and handles the changes of the complex background. Thus, more tasks can be performed by hand poses in contrast to finger tracking-based methods which result in a restricted gesture set in the interaction.

The flow diagram of the proposed method can be seen in the right side of Fig. 1. The camera grabs the projected background images only from a sub-region of the projected surface (recognition area). Out of the view of the camera the projected area can be used to display any information about the state of the recognition process (information area) e.g. pictogram of the detected gesture class. The first step is the foreground segmentation by our background subtraction method. This background image is updated when any change is detected in the projected image. The gesture module analyzes the segmented arm image and the result of the recognition gives the input of a Finite State Machine (FSM). The grammar of the FSM determines the actual task, which modifies the projected image. This task can be a command; e.g. order the system to step the presentation to the next slide or draw a line in the projected image. Because the projected image is known by the system the stored background image can be updated corresponding to the modified projected image. Finally, the projector displays the modified image and the processing cycle starts again. The next background subtraction is accomplished on the updated background image.

3.1 Segmentation Process

In the gesture recognition system the field of view of the camera is a subset of the projected region. Therefore any object in the projector beam reflects the exposure generating different texture patterns on the surface of the arm. For that reason the texture and the color of the hand is continuously changing according to the projected image and object position. These circumstances exclude any color segmentation or region-growing method for the segmentation. In that case the most popular solutions are based on finger tracking [4], but they restrict the usable gesture vocabulary. On that account we chose a background subtraction method and extended it to handle background changing. During projection the reflectance factor of the projected screen is near to 100% while the maximum for human skin is 70%, because the human skin partly absorbs the light, so it behaves as an optical filter [8]. Our method summarizes image difference with each image channel and foreground objects are classified by this summarized difference image by a threshold value. If the projector ray intensity is small at the position of the hand, e.g. the projected background is black, the difference between the hand and background reflection will be small and noisy. Hence the minimal projector lighting is increased above a threshold intensity value (in our case 20%) by linear histogram transformation of the projected image. The system only transforms the projected image during the interaction if any foreground object is detected by the segmentation method.

Since forearm features do not contain important information, the perfect and consequent segmentation of palm and forearm is important. The problem of automatic segmentation is introduced by other systems [9] [10]. We use a similar width-based wrist locating technique, which uses the main direction of the arm calculated from image moments. Considering this direction of the hand, the width of the wrist and that of the forearm can be measured. Analyzing width parameters of the forearm, the wrist position can be determined using anatomical structure information of the hand, because the calculated width values increase significantly at the wrist points from the forearm to the palm. Result of the segmentation process can be seen on Fig. 2.

Fig. 2. Steps of the contour segmentation; a) Input image; b) camera image with arm; c) segmented arm image; d) extracted palm contour

3.2 Calibration of the Camera-Projector System

The image grabbed from the camera involves the projected image in the background and any foreground object between the camera and the projected screen. In the camera image these objects are 2D projections of the 3D environment hence the contents of the image suffer from perspective distortions such as keystoning. Consequently, the system needs to register the coordinates of the pixels between the projected and its distorted version, which is grabbed by the camera. In the system this perspective distortion is modelled by a polynomial warping between the coordinates of the camera and the projector images. In our experiments the image warping of a first order polynomial was insufficient because higher order was required for the precise point registration. The second order polynomial equations can be expressed [7] as follows:

$$x' = a_0 + a_1 \cdot x + a_2 \cdot y + a_3 \cdot x^2 + a_4 \cdot xy + a_5 \cdot y^2 \qquad \textbf{(1)}$$
$$y' = b_0 + b_1 \cdot x + b_2 \cdot y + b_3 \cdot x^2 + b_4 \cdot xy + b_5 \cdot y^2$$

where (a_i,b_i) are the weighting coefficients of the geometrical warping, (x,y) the original and (x',y') are the new transformed positions. These input and output sample points are determined from the projection of a special calibration pattern image, or it can also be done by the edge content of images. Weighting coefficients are chosen to minimize the mean-square error between observed coordinate points and (x',y') coordinate points. After the geometrical calibration system may give the correspondence between the original projected and detected palm position.

3.3 Background Image Generation by A Priori Information

The main disadvantage of the background segmentation method is that it fails when background (user interface) significantly changes. In that case the well-known background updating techniques, e.g. running image averaging, does not work because certain regions of the projected image could alter behind the hand. Thus, we improved the segmentation method to overcome this problem by background image generation from the a priori information of the system configuration. However, the input of the projected background image is known, so we could generate an artificial background without any foreground object. In the segmentation process this updated image is used for the background subtraction. The main problems are that the camera-grabbed image suffers from color and geometrical distortion due to perspective projections, and the color transfer function of the camera and the projector. In the color calibration phase an intensity transfer function (look-up table - LUT) is generated from the intensity of the input image and the grabbed image. Sample intensity values are projected and grabbed by the system and these sample pairs (control points) are used for generating the LUT. Values between control points are best interpolated by a fifth order polynomial. The geometrical transformation

parameters are determined in the previous section and the image warping uses bilinear interpolation.

When the background changes the system warps the input image by geometrical warping equations, and then it is transformed by the calculated LUT to generate the correct background image for the image differencing. The system generates the background image when it detects any foreground object and the projected image changes significantly. This change detection is performed by a simple image differencing between consecutive projected image frames. Figure 3. demonstrates segmentation results by generated background image. This segmentation method gives satisfactory results only when used with the color and the geometrically transformed background image. During the interaction this initial background image will be refreshed by the original camera image for precise segmentation by running average method. If any new object appears in the projected image, it is detected by the background segmentation method. If more than one blob is detected, the system chooses the correct one by a labeling method. This labeling method chooses the correct arm object by tracking its last known position and size parameter in the previously segmented images. The labeled arm blob is used for the gesture analyzing. Regions of the residual detected blobs assign regions for the background refreshing. All assigned points of the background image are refreshed with the corresponding pixel from the camera image resulting a continuous background updating.

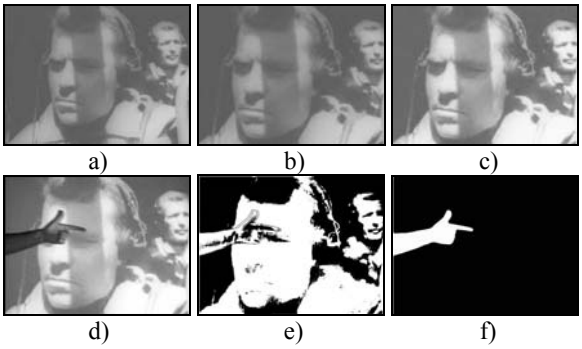

Fig. 3. Background segmentation results; a) the original projected image; b) geometrically transformed projected image; c) geometrical and color transformed projected image; d) real camera image with arm; e) background segmentation of the real camera image with geometrically transformed projected image; f) segmentation result on the geometrical and color transformed image (for color image versions see: http://almos.vein.hu/~licsara/projector)

4 Comparison of Feature Vectors for the Contour Classification

We applied a boundary-based method for the classification. Fourier descriptors are widely used for shape description e.g. character recognition [11], and in content-based image retrieval systems (CBIR) [14]. Recognition methods with Fourier

descriptors are usually based on neural networks classification algorithms [12][13] resulting 90-91% recognition rates for 6 gestures. In our method gesture contour is classified by nearest neighbor rule and the distance metric based on the modified Fourier descriptors [11] (MFD), what is invariant to transition, rotation and scaling of shapes. In these systems the examined shape should be defined by a feature vector, which is a closed curve, so the discrete function is periodic, to expand it into Fourier series. Our new feature vector approach can be seen on Fig. 4A. This feature is a complex coordinate vector (Method "A"), which is generated from the coordinate points of the palm boundary between wrist points as a complex sequence. The problem with that sequence is that it is not periodic, so we need to extend and duplicate it with its reversed sequence to get a periodic function. We compared the previous method with several feature extraction methods (Fig. 4B, C, D), which are widely used in CBIR systems [14]. The second method (Method "B") computes a similar boundary coordinate sequence, but it is generated from the contour of the whole hand mask. The next feature vector is derived from the centroid distance (Method "C") sequence, which is expressed by the distance of the boundary points from the centroid of the silhouette. Finally, the wrist centroid distance (Method "D") feature vector computes distances of the contour points from the centre of the wrist line.

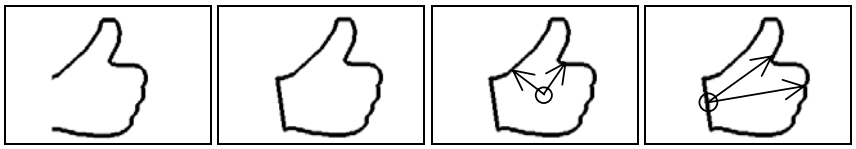

Fig. 4. Examined feature vectors for palm shape classification with the 4 different methods

The extracted feature sequence is classified by the modified Fourier descriptor. The method calculates the discrete Fourier transform (DFT) of this complex sequence. This method applies magnitude values of the DFT coefficients to be invariant to the rotation. We extended the MFD method to get symmetric distance computation. Denoting the DFT coefficients of the compared curves with F_n^1 and F_n^2, standard deviation function denoted by σ, the distance metric between two curves is as follows:

$$Dist(F_n^1, F_n^2) = \sigma\left(\frac{|F_n^1|}{|F_n^2|}\right) + \sigma\left(\frac{|F_n^2|}{|F_n^1|}\right) \quad \textbf{(2)}$$

We examined how many Fourier descriptors should be used in the distance computation. We measured the average efficiency of the recognition with several cut-off frequencies and feature extraction methods (Fig. 5). By determining the appropriate cut-off frequency the classification method is robust against noise of irregularities of the shape boundaries. One advantage of this robust method is that the

training set is very small. In our system the training phase is very fast because we store the average feature vectors of 20 consecutive gesture samples for each class.

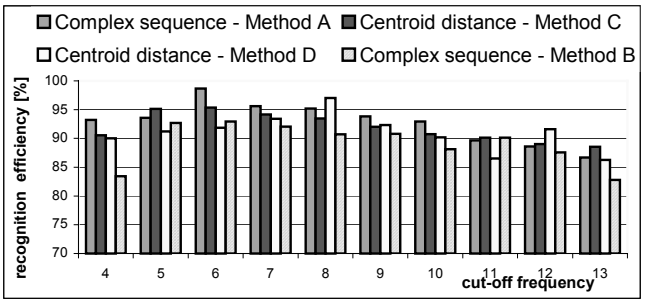

Fig. 5. Recognition efficiency by several cut-off frequencies

From the experiments we chose the first 6 coefficients (excluding the DC component) for methods "A", "C" and "D", and the first 8 coefficients for method "B". These results are utilized in our gesture recognition tests. Gestures of several users are tested with the proposed feature extraction methods. We have tested the recognition methods with 9 gesture classes and 400 gesture samples per person (Table 1). Each user trained all gesture classes before the recognition phase. Users can be found in the rows, while different feature extraction methods are in the columns.

Table 1. Pose classification results with several method and users

	Recognition rates [%]			
	Method A	Method B	Method C	Method D
Users				
User 1	99.6	96.8	98.1	97
User 2	98.3	92.6	96.7	98.1
User 3	97.9	95.2	95.3	91.2
User 4	98.2	91.2	94.5	92.3

It can be seen from our tests that the classification method gives better result, if the feature vector is calculated from the complex sequence of the boundary between wrist points (Method "A"). This approach gives more unambiguous features, since for example the shape contours of the palm when showing only the index or the thumb finger is very similar to each other, while the contour between wrist points are still distinct. This assumption is proved by the experimental results. The proposed system runs in simultaneous real-time performing image projection and grabbing tasks at resolution of 384*288 pixels on a single 1.7GHz Pentium processor.

5 Correction of the Recognition Efficiency

From our experiments we observed that during the interaction users perform gestures for a minimal time period (1-2 sec.). Therefore it can be supposed that results of the recognition should be stable for a given time except when the user changes the performed class. The distance is measured between the actual gesture and stored gesture classes with several consecutive gestures in time. On Fig. 6 the probability order of the faulty classified gesture classes is described. If the recognition of the actual gesture is false, the correct gesture class is not recognized as the most probable gesture (the measured distance corresponds to the probability of the recognized gesture).

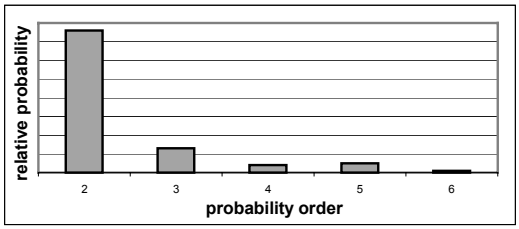

Fig. 6. Measured probability order of the unrecognized gesture classes

It can be seen that most of the misrecognized gestures were classified into the second most probable gesture class with a high probability. Consequently, we could derive gesture probabilities from several consecutive frames, and choose the most probable gesture class with maximum likelihood. If the occurrence of the unrecognized gestures is low then the detection of the misrecognized gestures can be improved. The estimated gesture class is detected by a maximum likelihood approach:

$$L = \arg\max_{i} \left(\prod_{t} \left(1 - d_{i,t} \right) \right) \tag{3}$$

Parameter $d_{i,t}$ is the measured distance between two gesture classes normalized by the maximum distance value into [0,1[interval. The zero distance means that the probability of the actual class is 1. Probability of the gesture recognition is calculated from the proposed distance metric, where parameter i identifies the gesture class, while t is the time parameter between consecutive frames. The estimated gesture (L) is determined by the maximum likelihood of gesture classes in the time domain. The standard gesture recognition (Section 4) is compared with the above maximum likelihood approach (Fig. 7). The test user performs 2 gestures with pose changing at frame #31. The recognition with standard method fails from frame #6 to frame #9. The proposed maximum likelihood (ML) estimation corrects this error in time. Between frames #31 and #33 the gesture recognition is unstable due to the gesture transition between two poses. The ML based estimation repairs this error and gives stable recognition result in time. In our experiments ML estimation was calculated in the time period of 6 consecutive frames. In Table 2 the recognition efficiency is summarized applying the standard and the maximum likelihood-based (ML) methods.

User tests involve consecutive gesture samples from each user. The ML based method corrects the misrecognized gesture classes and follows the gesture changing in time.

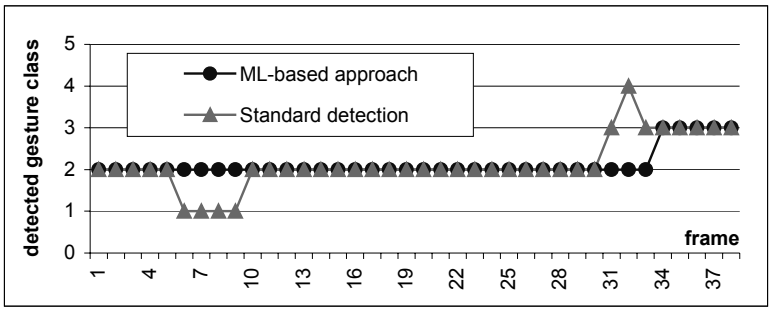

Fig. 7. Improving of the recognition efficiency by analyzing gestures in time

Table 2. Recognition results with standard and Maximum likelihood-based recognition

	Recognition rates [%]			
	User 1	User 2	User 3	User 4
Methods				
Standard method	99.6	98.3	97.9	98.2
ML-based method	100	99.3	99	99.6

6 Conclusions and Future Works

The vision-based gesture recognition system and the camera-projector configuration form a natural way to control multimedia presentations or manipulate directly the projected image. Our hand pose recognition system offers more freedom in communication for the speaker if compared to other methods using camera-projector systems. The above work has shown that the modified Fourier-based method is robust even with a small training set, so the modification of gesture vocabulary or retraining of gestures is more efficient. We have tested our feature extraction method against several methods from the literature and showed that our method is significantly efficient. We have shown that measuring the detection results by maximum likelihood approach in the time domain could significantly improved the recognition efficiency. Gesture-based systems are considered as typical user-independent recognition tools. Users would like to use them with high recognition efficiency without preliminary training of gestures. In our consecutive work we deal with interactive training of gestures to avoid retraining of all gestures if an untrained user would like to use the system.

References

1 Streitz, N.A., Geissler J., Haake J.M., Hol, J.: DOLPHIN: Integrated Meeting Support across Liveboards, Local and Remote Desktop Environments. Proceeding of the ACM CSCW, (1994) 345-358

2 Fraser, Q.S., Robinson, P.: BrightBoard: A Video-Augmented Environment. Proceedings of CHI'96. Vancouver (1996) 134-141

3 Wellner, P.: Interacting with Paper on the DigitalDesk. Communications of the ACM. (1993)

4 Hardenberg, C., Berard, F.: Bare-Hand Human Computer Interaction. Proc. of ACM PUI, Orlando (2001)

5 Hall, D., Gal, C., Martin, J., Chomat, O., Crowley, J.L.: MagicBoard: A contribution to an intelligent office environment. Robotics and Autonomous Systems 35. (2001) 211–220

6 Leubner, C., Brockmann, C., Müller, H.: Computer-vision-based Human Computer Interaction with a Back Projection Wall Using Arm Gestures. 27th Euromicro Conference. (2001)

7 Pratt, W.K.: Digital Image Processing, Wiley-Interscience, New York (2001)

8 Störring, M., Andersen, H. J., Granum, E.: Skin colour detection under changing lighting conditions. 7th Symposium on Intelligent Robotics Systems. Coimbra Portugal 20-23

9 Imagawa, K., Taniguchi, R., Arita, D., Matsuo, H., Lu, S., Igi, S.: Appearance-based Recognition of Hand Shapes for Sign Language in Low Resolution Image. Proceeding of 4th ACCV. (2000) 943-948

10 Koh, E.S.: Pose Recognition System. BE Thesis. National University of Singapore (1996)

11 Rui, Y., She, A., Huang, T.S.: A Modified Fourier Descriptor for Shape Matching in MARS. Image Databases and Multimedia Search. (1998) 165-180

12 Ng, C.W., Ranganath, S.:Real-time gesture recognition system and application. Image and Vision Computing 20, (2002) 993-1007

13 Chen, F.S., Fu, C.M., Huang, C.L.: Hand Gesture Recognition Using a Real-Time Tracking Method and Hidden Markov Models. Image and Vision Computing 21, (2003) 745-758

14 Zhang, D., Lu, G.: A Comparative Study of Fourier Descriptors for Shape representation and Retrieval. ACCV2002. Melbourne Australia (2002)