Dynamic Training of Hand Gesture Recognition 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

We developed an augmented reality tool for vision-based hand gesture recognition in a camera-projector system. Our recognition method uses modified Fourier descriptors for the classification of static hand gestures. Hand segmentation is based on a background subtraction method, which is improved to handle background changes. Most of the recognition methods are trained and tested by the same service-person, and training phase occurs only preceding the interaction. However, there are numerous situations when several untrained users would like to use gestures for the interaction. In our new practical approach the correction of faulty detected gestures is done during the recognition itself. Our main result is the quick on-line adaptation to the gestures of a new user to achieve user-independent gesture recognition.

1. Introduction

In this paper we demonstrate an effective human-computer interface for a virtual mouse system in a camera-projector configuration. 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 is to be performed near the computer. 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 presentation by hand gestures in a camera-projector system. For this purpose we use the image acquired by a camera observing gestures of the presentation in front of the projected image.

Gesture-based systems are considered as typical user-independent recognition tools. In our new practical approach, training and recognition are done in an interactive supervised way. Gesture parameters of previous interactions are applied as initial recognition parameters and the actual user continuously modifies these parameters in a supervised or unsupervised way. Supervised training means that the user retrains the wrong detected gestures while unsupervised recognition means continuous refreshing of detected gesture parameters during the recognition phase. The advantages of this approach are that only faulty detected gestures are retrained and the system is able to adapt to the gestures of new users without preliminary training. This implicit training method is fast and efficient as proved in our test. The proposed method has been tested by numerous users without any special knowledge about how the system works.

The main differences compared to the previous work [1] are that in the proposed system the gesture classification method is improved and the interactive training method is significantly reworked for the more efficient and comfortable employment. The retraining of all gestures or changing gesture vocabulary is also introduced as a significantly improvement in the proposed system. Moreover, the method is tested in a projector-camera system as a real-life application.

In the following we present related systems and give an overview of our work. In next sections the camera-projector calibration and hand segmentation methods are also introduced. Section 5 contains the gesture classification method itself. Finally, we describe the dynamic training method and test results with several subjects.

2. System overview and related works

A projector-camera pair is used to display the user interface on the projected surface where the camera acquires the projected image and the gestures of the user provide feedback about the interaction. System applies boundary-based method to recognize pose of static hand gestures. The virtual user-interface can be displayed on the projected background image and it is controlled by the detected hand gestures and palm positions. The aim is that the user should be able to interact with the projected

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

image instead of applying computer interfaces indirectly. In SmartBoard [1] there are special display hardware devices with sensors to detect physical contact with the display. BrightBoard [3] system uses a video camera and audio feedback to control the computer through painting simple marks onto the board. Other methods DigitalDesk [4], FreeHandPresent [5] apply hand gestures e.g. to navigate in a projected presentation by a restricted gesture set by counting and tracking fingers against 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. Method applies finger resting on an item for 0.5 seconds as a "pick up" gesture.

In our proposed method the arm and the forearm are segmented from the projected background and it recognizes hand shapes and not only fingers. Our method uses a large gesture vocabulary with 9 hand poses and handles the projected complex background. Camera grabs the projected background images only from the sub-region of the projected surface (recognition area). Out of the view of the camera the projected surface can be used to display any information about the state of the recognition (information area) e.g. pictogram of the detected gesture.

3. Calibration of the projector-camera 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 second order polynomial warping between the coordinates of the camera and the projector images. These polynomial equations can be expressed as follows [6]:

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

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. Weighting coefficients are chosen to minimize the mean-square error between observed coordinate points extracted from the camera image and *(x',y')* coordinate points.

4. Arm segmentation

In our gesture recognition system the camera's field of view is the subset of the projected region. Therefore any object in the projector beam reflects the exposure generating different texture patterns on the arm's surface. For that reason the texture and color of the hand continuously changes due to the projected image and object position. These circumstances exclude any color segmentation or region-growing method for the segmentation. In that case most popular solutions are based on finger tracking [5] but it restricts the usable gesture vocabulary. On that account we choose a background subtraction method and extend 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 [7]. 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.

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 [8] [9]. We use a similar width-based wrist locating technique, which uses the main direction of the arm calculated from the image moments. Considering this direction of the hand, the width of the wrist and the forearm can be measured. Analyzing width parameters of the forearm, the wrist position can be determined using anatomy structure of the hand, because the calculated width values increase significantly at the wrist points from the forearm to the palm (Figure 1).

Another problem is that the background segmentation is sensible to the changing of the projected background (user interface). However the input of the projected background image is known so we could create an artificial generated background. The main problems are that the camera-grabbed image suffers from color and geometrical distortion due to the perspective projections, and the color transfer function of the camera and the projector. In the color calibration phase a look-up table (LUT) is generated from the colors of the input image and that of the grabbed image. For the LUT generation a fifth order polynomial is fitted to the sample values. When the background changes the system warps the input image by geometrical warping equations (see Section 3.) and then it is transformed by the calculated LUT to generate a correct background image for the image differencing. During the interaction this initial background image will be refreshed by the original camera image for the precise segmentation.

5. Contour classification

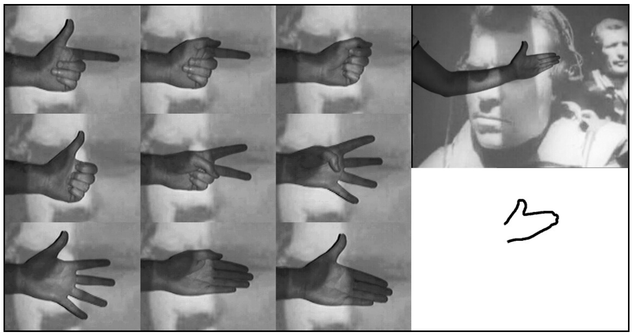

Figure 1. Gesture vocabulary and segmentation result

We applied a boundary-based method for the classification. Fourier descriptors are widely used for shape description (e.g. character recognition [10], and in content-based image retrieval systems. Recognition with Fourier descriptors is usually based on neural networks classification algorithms [11][12] resulting 90-91% recognition rate for 6 gestures. In our method the gesture contour is classified by the nearest neighbor rule and the distance metric based on the modified Fourier descriptors [10] (MFD). This metric is invariant to the rotation, transition, reflection, scaling of shapes. The examined shape should be defined by a feature vector, which is periodic, to expand it into Fourier series. Our method generates a feature sequence between the two wrist points (see Figure 1.) along the shape boundary. 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. The defined boundary sequence is constructed as a complex sequence of the x and y coordinates of the boundary points. The method then calculates the discrete Fourier transform (DFT) of this complex sequence. 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:

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

Our proposed method calculates the first 6 coefficients (excluding the DC component) therefore it is robust against noise of irregularities of shape boundaries.

We have tested the recognition method with 9 gesture classes (Figure 1). The starting set of training was very small, usually only one. This feature is important if we would like to use it for on-line training. In Table 1 the recognition efficiency of our pose classification method with several trainer and tester users can be seen. The trainer users can be found at the columns and the tester users are at the rows. It is tested by a set of 400 samples per user.

	Recognition results [%] Trainer users			
	User A	*User B*	*User C*	*User D*
Tester users				
User A	**99.8**	96.1	86.2	94.6
User B	90.4	**97.6**	90.1	93.9
User C	94.5	92.7	**99.6**	98.9
User D	95.5	95	96.7	**99.1**

Table 1. Pose classification results

If the trainer and tester user is the same person the recognition rates are above 97%, otherwise the results are above 86 %. In our experiments we observed that not all gestures' efficiency decrease significantly. Hence it would be more efficient if the system could learn only the faulty detected gestures by interaction with the user. The proposed system runs in simultaneous real-time at resolution of 384*288 pixels on a single 1.7GHz Pentium processor.

6. Dynamic training by supervised training

The goal of this approach is to avoid retraining of all gestures by new users. User corrects only the faulty detected gestures to improve the recognition efficiency of the system. The training method involves unsupervised (a) and supervised training (b). User follows the result of the recognition and he generates a user feedback by a special dynamic gesture if the result is not correct.

a) If the decision is right under the recognition phase (no user feedback), the detected gesture parameters will be continuously refreshed by the parameters of the current gesture. By continuously refreshing of the stored gesture parameters, the system is able to adapt to the small changes of hand poses automatically (unsupervised training). This training applies running average calculation between system parameters and actual gesture parameters. For example, when the user is tired and cannot show standard gestures, the system may learn it.

b) If the detection is incorrect, the user could notice that situation on the projected user interface and indicates the starting of the supervised training by performing the feedback gesture. This feedback gesture means fast and oscillatory moving of the hand like a handshaking. This feedback signal is user-independent since it uses only the velocity of the palm center and disregards the other results of the detection. We observed that this rapid hand shaking movement is infrequent during the interaction. When the variations of the palm position are greater than a threshold value during a given time (1-2 sec), the system recognizes a feedback signal. If the hand movement is fast, this summarized value will be greater than a predefined threshold, and the feedback signal will be detected. Threshold is adjusted in a way that more than one fast motion could produce detection to avoid false alarms. The supervised training has the following rules:

1. During the gesture training users need time to perform the displayed gesture correctly. The system records the sample for the training only if the detection of the performed gestures is consistent in time. Consequently the measured distance by MFD is stable for 3 seconds between consecutive hand contours.
2. During the training the system displays the pictogram of the selected gesture and projects homogenous white background for the perfect segmentation and training. This pictogram is segmented from the last training process and it is displayed out of the camera's view by the projector resulting continuous feedback about the recognition.
3. After the system accepts the gesture samples for the training, it advises and displays the next one due to the probability of gesture classes.
4. If the system advised the faulty detected class and the user corrected it, the training can be finished by the feedback signal. If the offered gesture is not the appropriate gesture, the user has to train the advised gestures until the correct gesture is displayed. At worst case all gestures will be retrained.

User order	Without dynamic correction	After dynamic gesture correction [%]	Number of correction steps
User A	99.5	-	-
User B	89.8	96,1	4
User C	94.7	97	2
User D	99.1	-	-
User B	93	98.8	1
User C	96.8	99.2	2

Table 2. Recognition results with several users applying dynamic training method

In our prototype application users can control a mouse-based prototype system for testing purposes. We defined several tasks for the users in which they should apply all 9 gestures. Users try the system successively in random order and test the recognition efficiency with and without dynamic training. The first user trains initial gesture parameters of the system and other users do not apply preliminary training. In Table 2 there are test results after dynamic gesture correction comparing to the normal recognition results without interactive training. At the last column the number of correction steps can be seen if user indicated gesture correction.

7. Conclusion

The above work has shown that dynamic training is user-friendly and user-independent. Gestures' classes can be trained from a limited number of training sets (it even works with solo training set) and supervised training is possible to correct faulty detected gesture classes. We have tested the supervised training system with several users and found that the performance of recognition has increased significantly as experimental data showed above. Our recognition results are above 96% that is more efficient than other Fourier-based methods.

8. References

[1] A. Licsár, T. Szirányi, "Hand Gesture Based Film Restoration", Proc. of PRIS'02, Alicante, 2002, pp. 95-103.

[2] N.A. Streitz, j. Geissler, J.M. Haake, and J. Hol, „DOLPHIN: Integrated Meeting Support across Liveboards, Local and Remote Desktop Environments", Proc. of the ACM CSCW, 1994, pp. 345-358.

[3] Q.S. Fraser, and P. Robinson, "BrightBoard: A Video-Augmented Environment", Proc. of CHI'96, 1996, pp. 134-141.

[4] P. Wellner, "Interacting with Paper on the DigitalDesk", Communications of the ACM, 1993.

[5] C. Hardenberg, and F. Berard, "Bare-Hand Human Computer Interaction", Proc. of ACM PUI, Orlando, 2001.

[6] W.K. Pratt, *Digital Image Processing*, Wiley-Interscience, New York, 2001.

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

[8] K. Imagawa, R. Taniguchi, D. Arita, H. Matsuo, S. Lu, S. Igi, "Appearance-based Recognition of Hand Shapes for Sign Language in Low Resolution Image", Proc. of 4th ACCV, 2000, pp. 943-948.

[9] E.S. Koh,. Pose Recognition System. BE Thesis, National University of Singapore, 1996.

[10] Y. Rui, A. She, T.S. Huang, "A Modified Fourier Descriptor for Shape Matching in MARS", Image Databases and Multimedia Search, 1998, pp. 165-180.

[11] C.W. Ng, and S. Ranganath, "Real-time gesture recognition system and application", *Image and Vision Computing* 20, 2002, pp. 993-1007.

[12] F.S.Chen, C.M. Fu, and C.L. Huang, "Hand Gesture Recognition Using a Real-Time Tracking Method and Hidden Markov Models", *Image and Vision Computing* 21, 2003, pp. 745-758.