15—1

Online Gesture Recognition Using Predicative Statistical Feature Extraction and Multivariate Analysis

Bisser Raytchev [1&2] , Osamu Hasegawa [2] , Nobuyuki Otsu [1&2]
[1] Department of Informatics & Electronics, Tsukuba University, Japan
[2] Machine Understanding Division, ETL, 1-1-4 Umezono, Tsukuba, Japan 305
Tel: 81-298-54-5080 (*) 71307, Fax: 81-(0)298-54-3313
E-mail:{bisser,hasegawa,otsu@etl.go.jp}

Abstract

A new method for gesture/motion recognition from time-varying image sequences is proposed, using predicative statistical feature extraction combined with linear discriminant analysis. The method offers natural, efficient and robust extraction/representation of information about motion and is at the same time computationally inexpensive. Good generalization abilities for gesture recognition are achieved by the method we propose: it is robust to changes in background and illumination conditions, to subjects' external appearance (clothing, body size, etc.), successfully copes with the non-uniformity in the speed of the gestures. No manual segmentation of any kind, or use of markers, sensors, etc. are necessary. The method requires no special environmental conditions, e.g. it is suitable for use in normal office environment. Real-time speed of processing can be achieved. Having these features, the method could be successfully applied to the creation of more refined human-computer interfaces. Also, since no domain-specific knowledge is used, the method can be easily adapted to other problems involving motion recognition.

1 Introduction

Recent increase in both computational power and storage capacity of personal computers, together with the availability of image acquisition devices at reasonable prices, have led to an increased interest in the creation of systems capable to provide more refined human-computer interaction (HCI) (see [1] for a recent review on the use of hand gestures for HCI, [2] for a survey on motion-based recognition research). Given the importance of visual information for us humans, gesture recognition will necessarily be a major component of such interfaces. For a successful gesture recognition system, good generalization abilities are essential, and for this end it has to be provided with the following features: to be robust to changes in background and illumination conditions; independence to subjects' external appearance (including gender, body size, clothing, etc.); ability to cope with the non-uniformity in the speed of the gestures (time invariance); to be robust to subjects' changes in position in space, both in the horizontal plane and in depth. Also, if the system is to be used as a part of a human-computer interface, real time performance is indispensable.

In this paper we propose a method for gesture recognition from time-varying image sequences, which utilizes predicative statistical feature extraction combined with linear discriminant analysis (LDA) for its discrimination/classification part. To evaluate the performance of the method, until now it has been tested with several different data sets, some of which have been created to incorporate some of the requirements for generalization abilities mentioned above. The method will be explained in some detail in section 2, followed by a description of the test experiments conducted and a report of the test results in section 3.

2 Description of the method

The method operates in two stages. At the first stage (primitive feature extraction), a set of primitive geometrical features related to motion changes are extracted from the input time sequence of moving images. Input data are mapped from pattern space $\boldsymbol{P}$ to primitive-feature space $\boldsymbol{F}$. At the second (learning) stage, the primitive features extracted at the first stage are linearly combined on the basis of multivariate analysis, using LDA [3,4], to provide new and more effective features (discriminant-feature formation). This learning process determines the mapping from primitive-feature space into discriminant-feature space ***D***, where different classes of gestures depict different trajectories (see Fig.1). Recognition is performed by comparing a test-sample gesture's trajectory in ***D*** to the trajectories of all classes of previously learned gestures and classifying it to that class which is most similar to it .

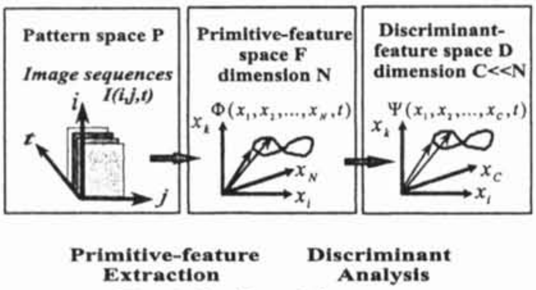
Pattern space P
Image sequences
I(i,j,t)
i
t
j
Primitive-feature space F dimension N
Φ(x1,x2,...,xN,t)
xk
xN
xi
Discriminant-feature space D dimension C<<N
Ψ(x1,x2,...,xC,t)
xk
xC
xi
Primitive-feature Extraction
Discriminant Analysis

Fig. 1: Outline of the system

What kind of primitive features would be most natural and effective for the representation of motion is the essential question which we have tried to tackle in this study. Our approach is to form short primitive geometric predicates, which are statistically integrated to obtain information about the quantity and type of motion change in consecutive frames. In practice, let *I(i,j,t-1)*, *I(i,j,t)* and *I(i,j,t+1)* be three consecutive image frames, where *i* and *j* are space coordinates and *t* is time. In each of these three frames we fix corresponding *reference pixels p(i,j,t-1), p(i,j,t)* and *p(i,j,t+1).* The predicate

$$P(\,T(\,(p(i,j,t)-p(i,j,t-1))^2\,)\;AND\;\; T(\,(p(i,j,t)-p(i,j,t+1))^2\,)\,) \qquad (1)$$

is formed, where the function *T (x)* is defined as:

$$T(x)=\begin{cases}1: & x\geq a\\ 0: & x<a\end{cases} \qquad (2)$$

and ***a*** is a threshold parameter. Depending on the values of the reference pixels, 4 different cases are possible: *1) P(0 AND 0); 2) P(0 AND 1); 3) P(1 AND 0); 4) P(1 AND 1).* "0" value reflects the fact that there has been no change in the corresponding reference locations, while "1" means that some change has occurred. These 4 cases divide primitive-feature space into 4 subspaces. Also, the following functions are calculated :

$$\mathrm{F}_k(t)=\sum_i\sum_j T((p_k(i,j,t)-p_k(i,j,t-1))^2);$$
$$\mathrm{G}_k(t)=\sum_i\sum_j T((p_k(i,j,t)-p_k(i,j,t+1))^2);$$
$$(3)$$
$$\mathrm{H}_m(t)=\sum_i\sum_j T((p_m(i,j,t)-p_m(i,j,t-1))^2)+ T((p_{m+4}(i,j,t)-p_{m+4}(i,j,t+1))^2);$$
$$\mathrm{H}_{m+4}(t)=\sum_i\sum_j T((p_{m+4}(i,j,t)-p_{m+4}(i,j,t-1))^2) +T((p_m(i,j,t)-p_m(i,j,t+1))^2);$$

k: 1..8, *m*: 1..4;

where T(x) is the threshold function (2), and pixel-value functions $p_s(i,j,t)$ are defined as follows:

$$\begin{aligned}
&p_1(i,j,t)=p(i+l,j,t)\,;\\
&p_2(i,j,t)=p(i+l,j-l,t)\,;\\
&p_3(i,j,t)=p(i,j-l,t)\,;\\
&p_4(i,j,t)=p(i-l,j-l,t)\,;\\
&p_5(i,j,t)=p(i-l,j,t)\,;\\
&p_6(i,j,t)=p(i-l,j+l,t)\,;\\
&p_7(i,j,t)=p(i,j+l,t)\,;\\
&p_8(i,j,t)=p(i+l,j+l,t)\,.
\end{aligned} \qquad (4)$$

In (4), *l* is a suitable parameter (distance between neighboring pixels) and *t* (time) can take one of the three values *t-1*, *t* or *t+1*, thus indicating from which of the three consecutive image frames the pixel function $p_s(i,j,t)$ is calculated. Each one of the functions in (3) represents a separate dimension in each of the four subspaces of feature space determined by (1), so that from (3) we have 26 different dimensions, which multiplied by four (for each subspace) gives 104 dimensions in all for primitive-feature space. In such way, during the learning process, for each training sample from each gesture class, the motion which has occurred in the time period between *t-1* and *t+1* is represented as one value of the feature vector Φ (*r,s,t*), where *r* (1 ... number of classes, i.e. different gestures in our case), *s* (1 ... number of training samples for each class) and *t* (1 ... number of frames for each training sample) successively take all possible values as the feature extraction is carried out. As time is changing, Φ (*r,s,t*) depicts a trajectory in primitive-feature space (see Fig.1). The feature functions (3) are projected by LDA in the lower dimensional discriminant-feature space

D, where a trajectory Ψ (r,s,t) corresponds to Φ (r,s,t). There, for each gesture class, a class-average trajectory is made from the trajectories of each training sample. Online recognition of a given test sample is performed by comparing the distance of its trajectory to each one of the class-average trajectories and classifying the test sample to that class to whose trajectory the distance is minimal (see Fig. 2).

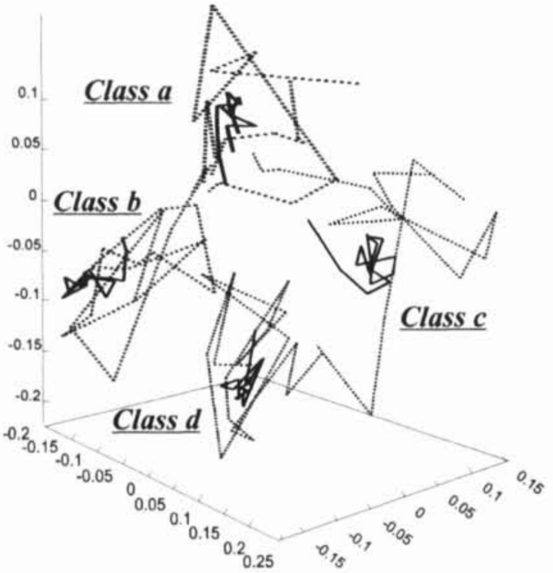

Fig.2: Gesture trajectories in discriminant-feature space for 4 classes of gestures. Class averages for each class are displayed by unbroken lines, while trajectories of the corresponding test samples (which have not been used during the training process) are shown in dotted lines.

3 Experimental results

To check the performance of the proposed method, it has been tested with the following different data sets.

3.1 Multimodal database of gestures with speech (MMDB) (available from [5]). This database contains time-varying images of gestures of upper body. The following 9 classes of gestures were used for the test: 1) up (lift one hand up); 2) right (move one hand to the right); 3) left (move one hand to the left); 4) me (pointing to oneself); 5) right circle; 6) left circle; 7) stop; 8) expand; 9) reduce. Several snapshots from these image sequences are shown in Fig. 3. Data from 6 different subjects (3 women and 3 men) with 4 samples from each gesture were used. The recognition rates were estimated by the leave-one-out method and an average recognition rate of 95.4 % was reached.

Since the gestures in the MMDB database have been taken under some restrictions (e.g. uniform background, clothing, etc.), it was necessary to check the method with data taken under more "real-world" conditions. All of the following data has been taken in the conditions of an usual office environment, with no special illumination (fluorescent lights on the ceiling were the only light source) or other special conditions. 10 different classes of gestures were used: 1) bow; 2) move head saying "no"; 3) move right hand up; 4) move left hand up; 5) move left hand and upper body left; 6) move right hand and upper body right; 7) clap hands; 8) banzai (move both hands up); 9) make a cross with both hands; 10) no motion. The subjects performed the above gestures while sitting on a chair in front of the video camera at a distance far enough so that all gestures could be captured (the distance and camera angle have not been fixed). The gestures were performed at a speed the subjects liked, i.e. they didn't have to imitate a certain fixed speed of performance. The following 3 experiments were conducted:

3.2 One subject performs the above 10-class gestures in very different clothing. 10 samples were taken from each gesture in 5 different clothes. Recognition rate better than 95% was reached.

3.3 Different subjects (4 men and 2 women) perform the above 10-class gestures in different clothing (see Fig. 4 for several snapshots of some of the image sequences used). For each person 2 samples were taken from each gesture in 2 different clothes. Average recognition rate better than 85% was reached on a leave-one-out test.

3.4 To check robustness to different background conditions, data has been made where one subject performs the above 10-class gestures on 3 very different backgrounds (under very different illumination conditions). Gestures on two of the backgrounds were learned, and tested on the third background (see Fig. 5 for several snapshots of some of the image sequences used). Recognition rate better than 87% was reached.

The training samples used in the last two data sets (3.3 and 3.4) were too few, and it is thought that more samples would significantly increase performance. In all cases mentioned above, real-time speed of processing can be achieved even on a personal computer with a fast enough processor (we used a PC with DEC Alpha 500 MHz processor).

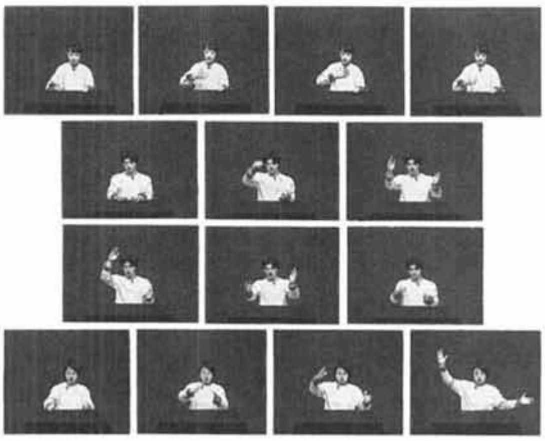

Fig. 3 Snapshots of some of the gestures used in Experiment 3.1 (MMDB data base)

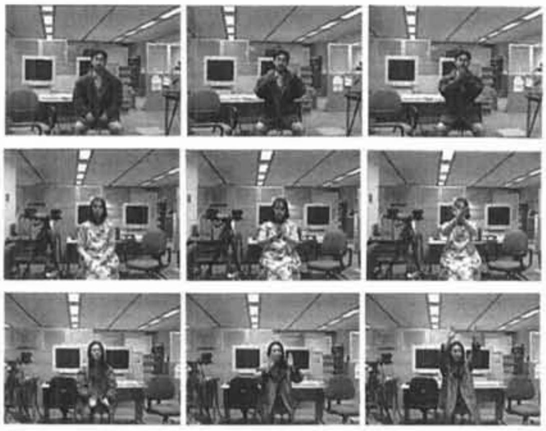

Fig. 4 Snapshots of some of the gestures used in Experiment 3.3 (different subjects).

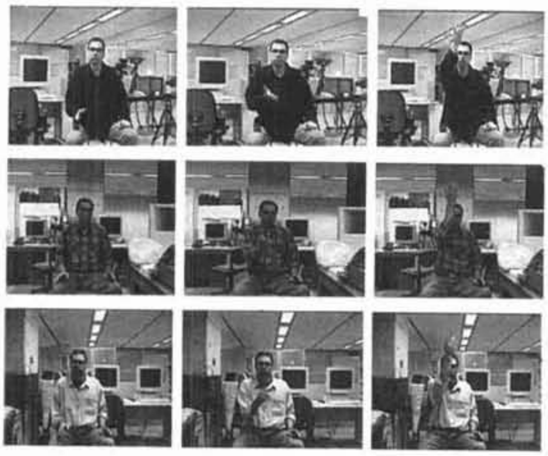

Fig. 5 Snapshots of some of the gestures used in Experiment 3.4 (different backgrounds and illumination conditions).

4 Conclusion

Even though being trained with only a very small number of gesture samples, the method we propose shows good generalization abilities (robust to changes in back-ground, illumination conditions, subjects' external appearance, non-uniformity in the performance speed of the gestures, etc.) and as such could be useful as a part of a user-independent human-computer interface. It is necessary, however, to further test the method more extensively with larger data sets, to determine the upper limit of its performance. Presently, the method works under the assumption of a single user at a time, and the possibility for multiple users performing different gestures at the same time should be investigated.

Another useful feature of this method is that since no domain knowledge is explicitly used, it could be easily adapted and applied to other problems involving motion recognition.

Acknowledgment

We would like to thank Dr. K. Sakaue from ETL for his help and encouragement in conducting this research. This research has been conducted in the frame of the Real World Computing (RWC) Project and we also would like to thank to all involved.

References

[1] V.Pavlovic, R. Sharma and T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review" , IEEE Trans. PAMI, Vol. 19, No. 7, 677- 694, 1997.

[2] C. Cedras and M. Shah, "Motion-based recognition : a survey", Image and Vision Computing, Vol. 13, No. 2, 129-155, 1995.

[3] T.W. Anderson, *"An Introduction to Multivariate Statistical Analysis"*, Wiley, New York, 1984.

[4] T. Kurita, N. Otsu and T. Sato, A Face Recognition method Using Higher Order Local Autocorrelation and Multivariate Analysis, Proc. ICPR, 213-216, 1992.

[5] S. Hayamizu et al.,: Multimodal Database of Gestures with Speech, Technical Report of IEICE,1996. Database available from:
http://www.rwcp.or.jp/wswg/rwcdb/mm