4P－2

Gesture Recognition By Geometrical Statistical Feature Extraction And Discriminant Analysis

Bisser Raytchev[1&2], Osamu Hasegawa[2], Nobuyuki Otsu[2]

[1] Department of Informatics & Electronics, Tsukuba University, Japan

[2] Machine Understanding Division, ETL, 1-1-4 Umezono, Tsukuba, Japan 305

{bisser, hasegawa, otsu}@etl.go.jp

1. Introduction

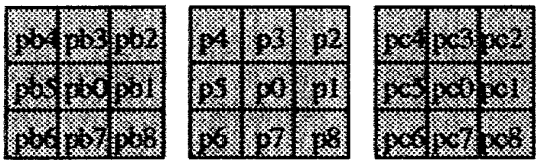
Recent increase in both computational power and storage capacity of personal computers, together with the availability of image acquisition devices at reasonable prices, have led to an increased interest in the creation of systems capable to provide more refined human-computer interaction. Given the importance of visual information for us humans, gesture recognition will necessarily be an important component of such interfaces. Real-time performance and good generalization abilities are essential in this case.

In this paper we propose a method for gesture recognition from time-varying image sequences, which utilizes geometrical statistical feature extraction combined with linear discriminant analysis for its discrimination/classification part. To evaluate the performance of the method, it has been tested with several different data sets (see section 3.).

2. Description of the method

The method we propose operates in two stages. During the first stage (feature extraction), from the input time sequence of moving images, a set of primitive geometrical features related to motion changes are extracted in a way which will be explained below. Each separate primitive feature represents a separate dimension in a feature space F, i.e. the process of feature extraction transforms the raw pixel data from pattern space P into data in feature space F, the latter being of much lower dimension, because most of the information not related to the task at hand has been eliminated. During the second stage (discrimination/classification), the data from feature space F are linearly combined on the basis of multivariate analysis[1] and further transformed into a category space C, which is characterized by a further dimensionality reduction and information compression. In category space, the images of the different input samples belonging to the same class are combined to form a class average's trajectory, which is to be used for comparisons in the recognition part. The above-mentioned two stages constitute the process of learning, during which the two mappings F (from pattern space P into feature space F) and C (from feature space F into category space C) are established. The process of recognition functions identically: the test sample data is mapped by F-matrix into feature space, and by C into category space, where it is classified by measuring how near/far it is from each class average's trajectory.

What kind of primitive features are necessary and sufficient for the effective representation of motion is the essential question which we have tried to tackle in this study. Our approach is to form short geometric predicates, answers to which would contain statistical information about the quantity of motion change in consecutive image frames. 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. We form masks $M(i,j,t-1)$, $M(i,j,t)$ and $M(i,j,t+1)$, which correspond to each of the above image frames (see Fig.1). Each mask contains 9 grayscale pixel values at the corresponding locations, neighboring pixels being separated from each other by the same distance. Image frame $I(i,j,t)$ is the reference frame, and mask $M(i,j,t)$ - reference mask. Pixels pb0, p0 and pc0 we call reference pixels. The predicate



$M(i,j,t-1)$ $M(i,j,t)$ $M(i,j,t+1)$

Fig.1 Masks used for the feature extraction

P(T($(p0-pb0)^2$) AND T($(p0-pc0)^2$)) is formed, where the function T is defined as:

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

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 feature space F into 4 subspaces, F1 ... F4. Also, the following functions are created, identically to (1):

$$
\begin{aligned}
& T_i^b\left((p_i - pb_i)^2\right) \qquad i: 0..8; \\
& T_i^c\left((p_i - pc_i)^2\right) \qquad i: 0..8; \\
(2) \qquad & T_i^{bc}\left((p_i - pb_i)^2 + (p_j - pc_j)^2\right) \\
& T_i^{cb}\left((p_j - pb_j)^2 + (p_i - pc_i)^2\right) \\
& \qquad i: 1..4, \quad j: 5..8;
\end{aligned}
$$

Each of these functions represents a separate dimension in each of the four subspaces of feature space, so that from the above functions we have 26 dimensions, which multiplied by four (for each subspace) gives 104 dimensions in all. The masks M are simultaneously shifted from left to right and from top to bottom in the corresponding image frames, and for each mask shift, data for each dimension is calculated and added on the axis for that dimension. In this manner, the motion which has occurred in the time period between *t-1* and *t+1* is represented as one value of the feature vector $F(k,s,t)$, where k, s and t represent respectively gesture class number, sample number in class, and frame number.

3. Experiments and results

To check the performance of the proposed method, until now it has been checked with the following data sets. Real-time performance has been achieved on a personal computer with DEC Alpha 500MHz processor.

3.1. Multimodal database of gestures with speech (MMDB)[2].This database contains time-varying images of gestures of upper body. 9 classes of gestures were used: 1) up (move one hand up); 2) right; 3) left; 4) me (pointing to oneself); 5) right circle; 6) left circle; 7) stop; 8) expand; 9) reduce. Data of 6 different subjects (3 women and 3 men) with 4 samples of 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 (lowest among them 91.7%).

3.2. 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 (in Fig.2 are shown two frames of the samples used).

Fig. 2 Snapshots of some of the gestures used.

10 different classes of gestures have been 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.

- Only one subject performs the above 10-class gestures in different clothing. 10 samples were taken from each gesture in 4 different clothes. Recognition rate better than 95% was reached.
- Different subjects (4 men and 2 women) perform the above 10-class gestures in different clothing. For each subject 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.
- 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. Recognition rate better than 87% was reached.

4. Conclusion

The method we propose shows very good generalization abilities (robust to changes in background, illumination conditions, subjects' appearance, non-uniformity in the performance speed of gestures, etc.) and is computationally inexpensive, allowing real-time performance to be achieved.

Acknowledgment We would like to thank Dr. K. Sakaue from the Adaptive Vision Lab ETL for providing the conditions for this research. This research has been conducted in the frame of the RWC Project and we also thank to all involved.

References

[1] T.Kurita, N.Otsu, and T.Sato: *A Face Recognition Method Using Higher Order Local Autocorrelations And Multivariate Analysis*, ICPR Proc., 1992

[2] 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.