CHAPTER 6

Experimental results

To check the performance of the system based on the methods proposed in the previous chapters, several different data sets have been used as input data. In all experiments mentioned in this chapter, the parameters from the previous chapters were fixed to proper constant values, i.e. no parameter tuning has been done in test mode. Real-time speed of processing has been achieved in all cases (except for the version using the ensemble of classifiers method) on a general purpose personal computer with DEC Alpha 500 MHz processor.

6.1. Multimodal database of gestures with speech (MMDB)

The multimodal database of gestures with speech (Hayamizu, 1996) has been created in the framework of the Real World Computing Project and is presently available from the following web site: www.rwcp.or.jp/wswg/rwcdb/mm. This database contains time-varying image sequences of gestures of upper body taken in front of a uniformly black

background. The duration of each gesture is about 3-4 seconds. The data has been saved as RGB files using an Indy SGI workstation. The images are in 320x240 format. The following 9 classes of gestures were used for the test (only the first 60 frames have been used): 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 (the image sequences can be seen at the WWW address given above). Only the first 60 frames from the available image sequence have been used. Data from 6 different subjects (3 women and 3 men) with 4 samples from each gesture were used (Fig.6-1). The recognition rates were estimated by the leave-one-out method, i.e. system was trained on data from five different subjects and tested on the sixth one. An average recognition rate of 96.8 % was reached.

Figure 6-1. Snapshots of some of the gestures used in Experiment 6.1 (MMDB data base). The following gestures are shown: "me" (*top*), "left circle" (*middle*) and "expand" (*bottom*).

6.2. "Real-world" data

To check the system with data taken under more "real-world" conditions (the gestures in the previous experiment were performed in somewhat artificial conditions – uniformly black background, same color and texture for the clothes, same distance to the camera, same speed of performance, etc.), we conducted some further experiments. All of the following data has been taken in the conditions of ordinary office environment, with no special illumination (fluorescent lights on the ceiling were the only light source) or any other special conditions (see Figs. 6-2, 6-3, 6-4). Ten 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" (lift both hands up); 9) "make a cross with both hands"; 10) "no motion". Several snapshots from some of these gestures are shown below. 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 natural speed, i.e. the subjects didn't have to imitate a certain fixed speed of performance. The gesture sequence was demonstrated only once to the subjects, and after that they were asked to perform the gestures in a manner which they find preferable (as a result, some gestures were performed in quite a different way depending on the subjects' prefernces). Data was input to the memory of a personal computer (where all the information processing has been done) using a SONY DCR-VX1000 digital camera. The frame size of the input images was 256x240 and the duration of each gesture about 2 seconds. The following experiments were conducted.

6.2.1. Single user

In this experimental setting, one subject performs the 10-class gestures described above, wearing clothes with different color and texture (see Fig. 6-2 for several snapshots from the gesture sequence). Ten samples were taken from each gesture in 6 different kinds of clothes. Gestures with 5 different clothes have been learnt, and the test was performed on the sixth one. Average recognition rate of 97.2% has been reached on the leave-one-out test.

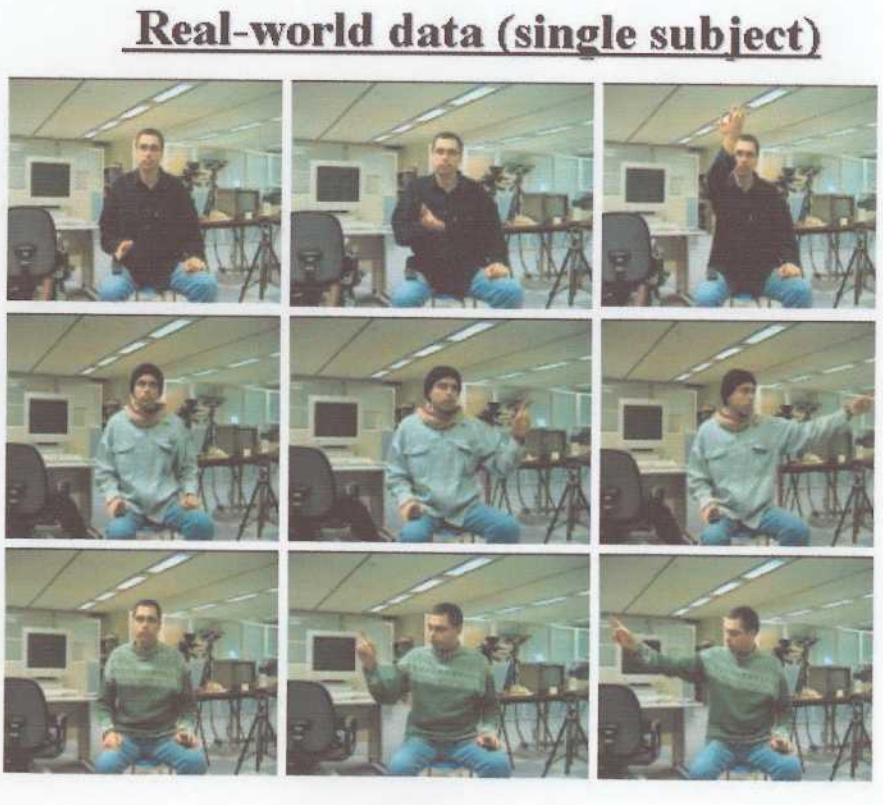

Figure 6-2. Snapshots of some of the gestures used in Experiment 6.2.1 (single user). The following gestures are shown: "move right hand up" (*top*), "move left hand to the left" (*middle*) and "move right hand to the right" (*bottom*).

6.2.2. Different subjects

In this experimental setting, different subjects (4 men and 2 women) perform the same 10-class gestures described at the beginning of section 6.2 (see Fig. 6-3 for several snapshots). For each person 3 samples were taken from each gesture in 2 different kinds of clothes. An average recognition rate of 90.3% was reached on the leave-one-out test (i.e. data from 5 subjects has been learnt, and tests done on the remaining 1 subject).

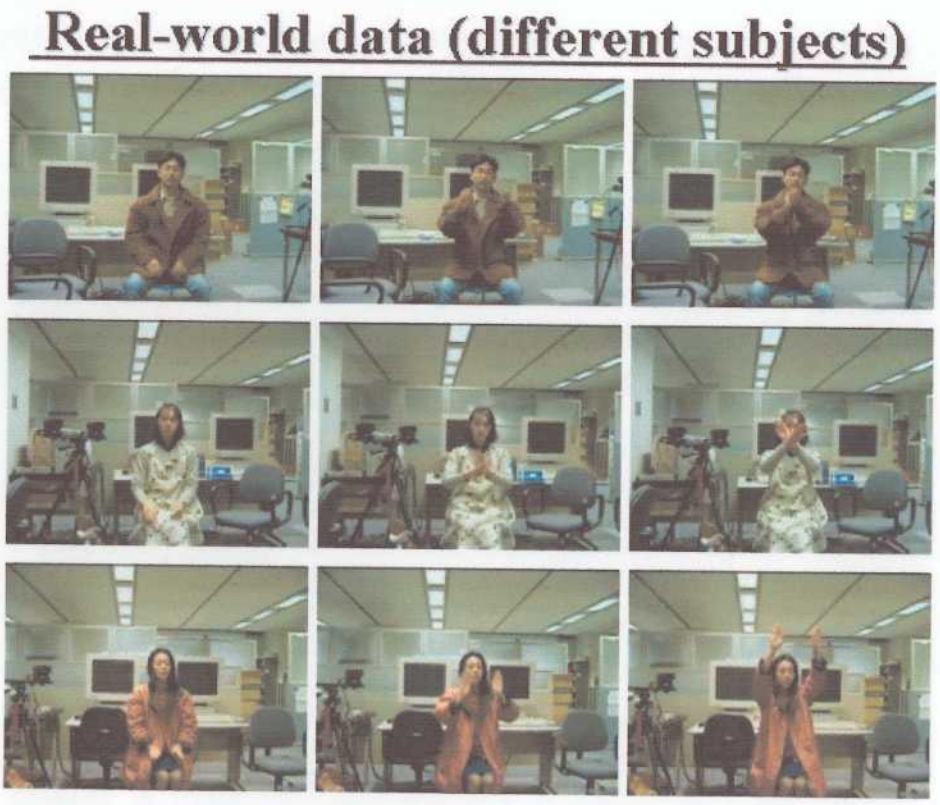

Figure 6-3. Snapshots of some of the gestures used in Experiment 6.2.2 (different subjects). The following gestures are shown: "clap hands" (*top*), "cross" (*middle*) and "banzai" (*bottom*).

The results obtained for this experiment are displayed in the table on Fig. 6-4, where each row represents the results for a different subject. The number of mistakes for each gesture class are shown, with the wrong gesture class following in the brackets. As can be seen from this table, for some subjects the recognition rate is as high as 100.0% (it should be noted that data from each subject evaluated here has not been used in the training sequence, i.e. the subject was totally unknown to the system), while in the worst case - as low as 75.0%. This can be explained with the fact that too few samples have been used in the training sequence, the data was quite noisy, and especially in the case of subject B. (for whom the performance was lowest) the distance from the camera was significantly different compared to the other subjects' cases, resulting in a significant difference in the size of the motion patterns (the size invariance normalization described in section 3.3 has not been used here).

Different subjects - test results (leave-one-out)

	A BOW	B TURN HEAD	C UP (RIGHT HAND)	L UP (LEFT HAND)	D LEFT	E RIGHT	G CLAP HANDS	I BANZAI	K CROSS HANDS	N NO MOTION	% CORRECT (error)
SUBJ. H (MALE)	2(K)	1(K)	0	0	0	0	0	0	3(G)	0	90.0% (6/60)
SUBJ. S (MALE)	0	1(E)	0	1(A) 1(I)	0	0	0	1(C)	0	0	93.3% (4/60)
SUBJ. W (MALE)	0	0	0	0	0	0	0	0	0	0	100% (0/60)
SUBJ. O (FEMALE)	1(L) 3(K)	0	0	0	0	0	0	2(K)	1(G)	0	88.3% (7/60)
SUBJ. A (FEMALE)	0	1(K)	0	0	0	0	0	2(L)	0	0	95.0% (3/60)
SUBJ. B (MALE)	1(G) 2(C)	3(C)	0	2(I) 1(C)	0	0	2(B)	0	3(E) 1(A)	0	75.0% (15/60)

Figure 6-4. Test results for the different subjects in experiment 6.2.3. Data shows number of mistakes (and mistaken class in the brackets) for 6 test samples

In Fig. 6-5 are shown the results from the same experimental data (6 different subjects), but in the case when the ensemble of classifiers algorithm presented in section 4-4 is used. The different columns show the recognition rates obtained for each classifier separately, and the last column shows the results obtained by the whole ensemble using the rule (4-17). The average recognition rate for the whole ensemble is about 3% higher than the best among the individual classifiers (the classifier for resolution 64x60 in this case). The mistaken gestures (shown here only for subject S.) are in the following format: e.g. A1(K) means that the 1st test sample for gesture A has been mistakenly classified as gesture K by the respective classifier. As can be seen the different classifiers' mistakes don't overlap in most of the cases which is used to achieve better performance by the majority voting rule. For example, I4 is seen as gesture C only by the 64x60 classifier, but all other classifiers vote for the correct class (class I), which is selected as the final classification result by the majority voting rule (4-17). Of course, if a certain sample is mistakenly classified by all or by the majority of the classifiers (e.g. gesture B1 is seen as class E by 3 of the classifiers and as class K by one), the resultant decision will be wrong as seen in Fig. 6-5. In the case when a majority cannot be reached, different strategies are possible. The test sample in this case can be rejected as not belonging to any of the learnt classes, or if it is known a priori (like in the present case) that only samples from the classes which have been learnt will be input, the output of the classifier with highest a posteriori probability (4-19) can be selected as the final decision. For example, in Fig. 6-5 for the subject S.'s case (shown with blue letters) two from the classifiers mistakenly see gesture G5 as gesture K and two classifiers see it correctly as gesture G. Because in this case the classifier with highest a posteriori probability is the 128x120 classifier, its output will determine the final decision, which will be correct in this case. Another example (not so lucky this time) is shown with black letters for test sample A1, which again is seen correctly as class A by two of the classifiers, and incorrectly as class K by the remaining two. However, in this latter case

the classifier with highest a posteriori probability is mistaken, and the output decision will have to be counted as an error.

Test results for an ensemble of classifiers at different resolutions (6 subjects; leave-one-out)

RESOLUTION / SUBJECT	256 x 240 (l=30,60)	128 x120 (l=15,30)	64 x 60 (l=10,20)	32 x 30 (l=5,10)	whole ensemble %CORRECT
SUBJECT H (MALE)	83.3%	85.0%	90.0%	85.0%	90.0%
SUBJECT S (MALE)	90.0% A1(K); B1(E); G2(K); G4(K); G5(K); K5(A)	91.7% A1(K); B1(E); B2(E);B5(E); G5(K);	93.3% B1(E);I4(C); L5(A); L6(I);	96.7% B1(K);K6(I);	96.7% B1: E(3) K(1); G5: K(2) G(2); A1: K(2) A(2);
SUBJECT W (MALE)	98.3%	98.3%	100.0%	100.0%	100.0%
SUBJECT O (FEMALE)	95.0%	88.3%	88.3%	91.7%	93.3%
SUBJECT A (FEMALE)	85.0%	90.0%	95.0%	90.0%	96.7%
SUBJECT B (MALE)	78.3%	71.7%	75.0%	73.3%	81.7%
AVERAGE	88.3%	87.5%	90.3%	89.5%	93.1%

Figure 6-5. Test results for the different subjects (experiment 6.2.2) using an ensemble of classifiers at different resolutions. Data shows percent correctly recognized from 60 test samples. The data for subject S. illustrates in more detail how better recognition rates can be achieved when using an ensemble of classifiers (see text for further explanation).

6.2.3. Different backgrounds and illumination conditions

To check robustness to different background and illumination conditions, data has been taken where one subject performs the above 10-class gestures on 3 different backgrounds under different illumination conditions. Three samples from each gesture were learned on two of the backgrounds, and tests were performed on the third background in a leave-one-out test (Fig. 6-6). Recognition rate better than 89% was reached.

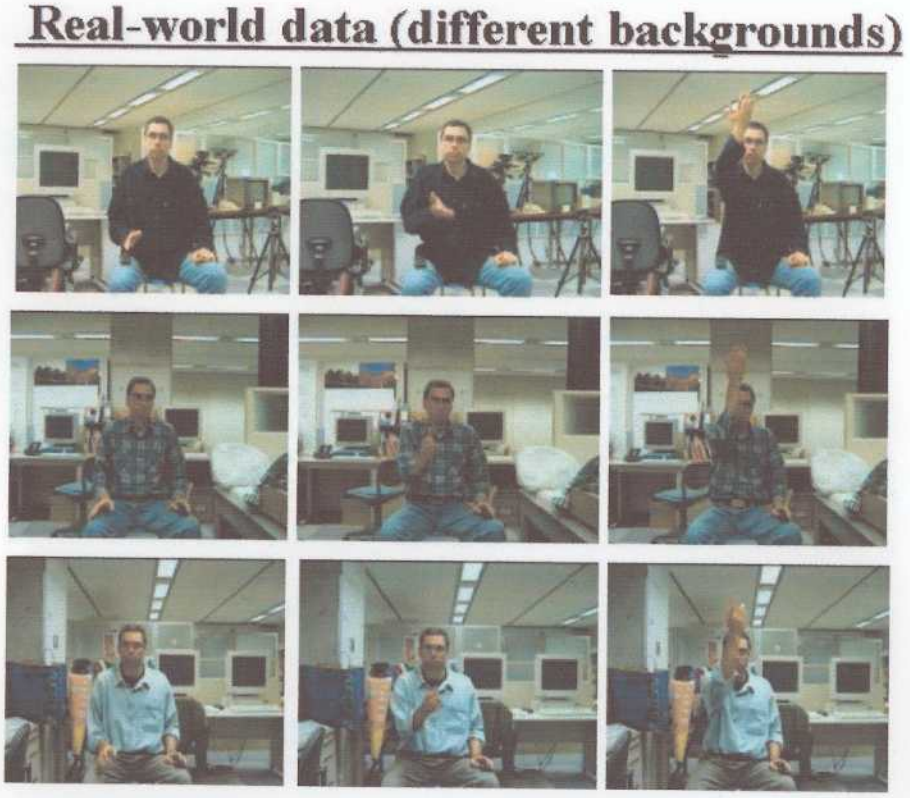

Figure 6-6. Snapshots of some of the gestures used in Experiment 6.2.3. The gesture "right hand up" is shown in three different background and illumination conditions.

The training samples used in this data set also were too few, and it is expected that using more samples would further increase performance.

6.2.4. Tests for size invariance

A separate data set has been created in order to check how system's behavior would be influenced when the size of the subjects performing the gestures changes significantly. In this experimental setting, different subjects (4 men and 2 women) perform the same 10-class gestures described at the beginning of section 6.2 (see Fig. 6-7 for several snapshots) but in different experimental environment (most of the subjects who participated in these experiments were also different from those in experiments 6.2.2). For

each person 5 samples were taken from each gesture at 2 different distances from the camera (at 5 and 7 meters respectively). The system was trained using the data taken at 5 meters distance, and after that tested on data taken at 7 meters. The feature normalization algorithm proposed in section 3.3 (formulae 3-13 – 3-16) has been used during the training stage.

Test for Size Invariance

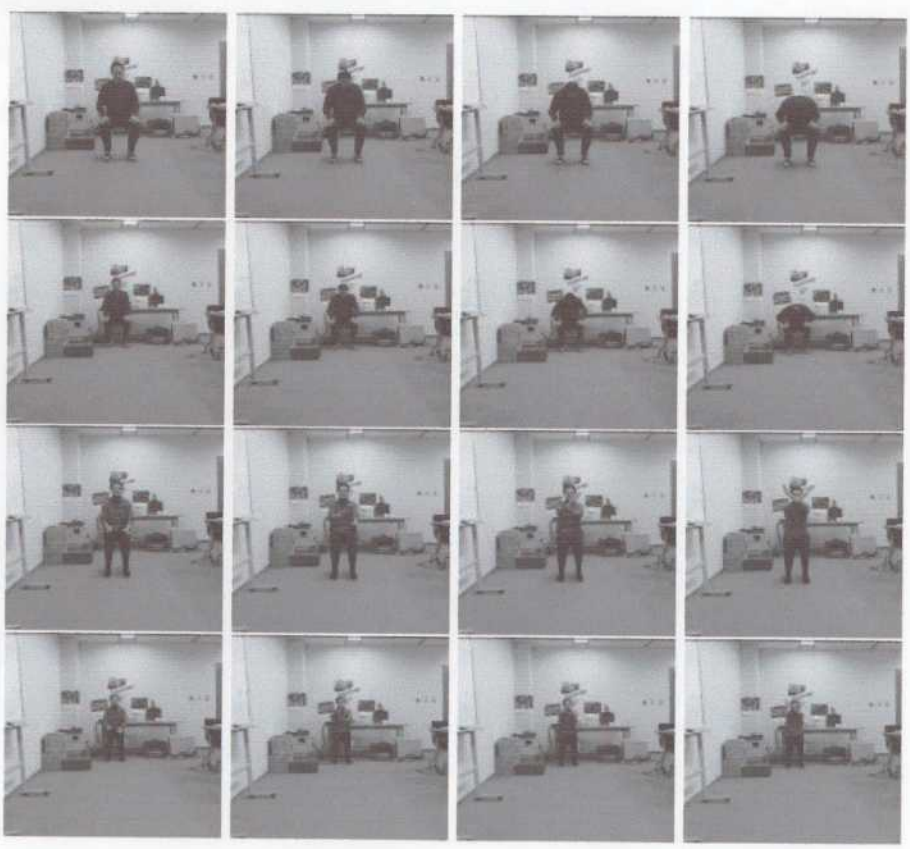

Figure 6-7. Snapshots of some of the gestures used in Experiment 6.2.4 (scale invariance tests). Several snapshots are shown for the gestures "bow" (first and second row) and "cross hands" (lower two rows). The gestures are learnt at distance 5 meters from the camera (first and third rows) and tests are done using input data taken at 7 meters from the camera (second and fourth rows).

In these tests, an average recognition rate of 81.7% was reached. The results obtained for this experiment are displayed in the table on Fig. 6-4, where each row represents the

results for a different subject. The number of mistakes for each gesture class are shown, with the wrong gesture class following in the brackets. As can be seen from this table, the system is capable to handle this situation, although somewhat lower recognition rates were obtained. The main reason for the relatively inferior performance in this case seems to be the fact that the redundancy of the information available from the binary motion patterns diminishes significantly as the distance from the camera increases and the area of the motion patterns (the useful area) becomes very small compared to the area of the whole image. This situation was further aggravated by the bad illumination conditions due to which the motion texture obtained from the differencing was too scarce compared with the previous experiments.

Test results for the size invariance test

GESTURE Subject (gender)	A BOW	B TURN HEAD	C UP (RIGHT HAND)	L UP (LEFT HAND)	D LEFT	E RIGHT	G CLAP HANDS	I BANZAI	K CROSS HANDS	N NO MOTION	% CORRECT (error)
B(M)	0	1(G) 1(N)	1(B) 1(L)	1(C)	0	0	1(A)	1(G)	2(G) 1(A)	0	80.0% (10/50)
X(M)	0	0	1(A) 1(I) 1(L)	0	0	1(G)	1(E)	1(B) 2(C)	2(G)	0	80.0% (10/50)
S(F)	0	2(G) 1(K)	1(L)	1(C)	0	0	1(C)	0	3(A) 1(G)	0	80.0% (10/50)
K(M)	2(B)	0	1(E) 1(L)	0	0	0	0	1(B) 1(G)	3(B)	0	82.0% (9/50)
O(M)	0	1(A)	1(K)	2(C)	0	0	0	1(E)	2(G) 2(B)	0	82.0% (9/50)
D(F)	0	1(G)	1(B)	2(C)	0	0	0	0	2(A) 1(B)	0	86.0% (7/50)

Figure 6-8. Test results for the different subjects in experiment 6.2.4. Data shows number of mistakes (and mistaken class in the brackets) for 5 test samples.

6.2.5. Tests under some more extreme conditions

The system's performance was further tested under some more extreme conditions (see Fig.6-9) for several snapshots from these tests:

(a) background and illumination conditions were chosen to be very different from those under which learning was performed and video camera's angle and position were also significantly changed;

(b) some other motion, not related to the task at hand, was introduced in the background to see how such type of noise would interfere with the performance of the system.

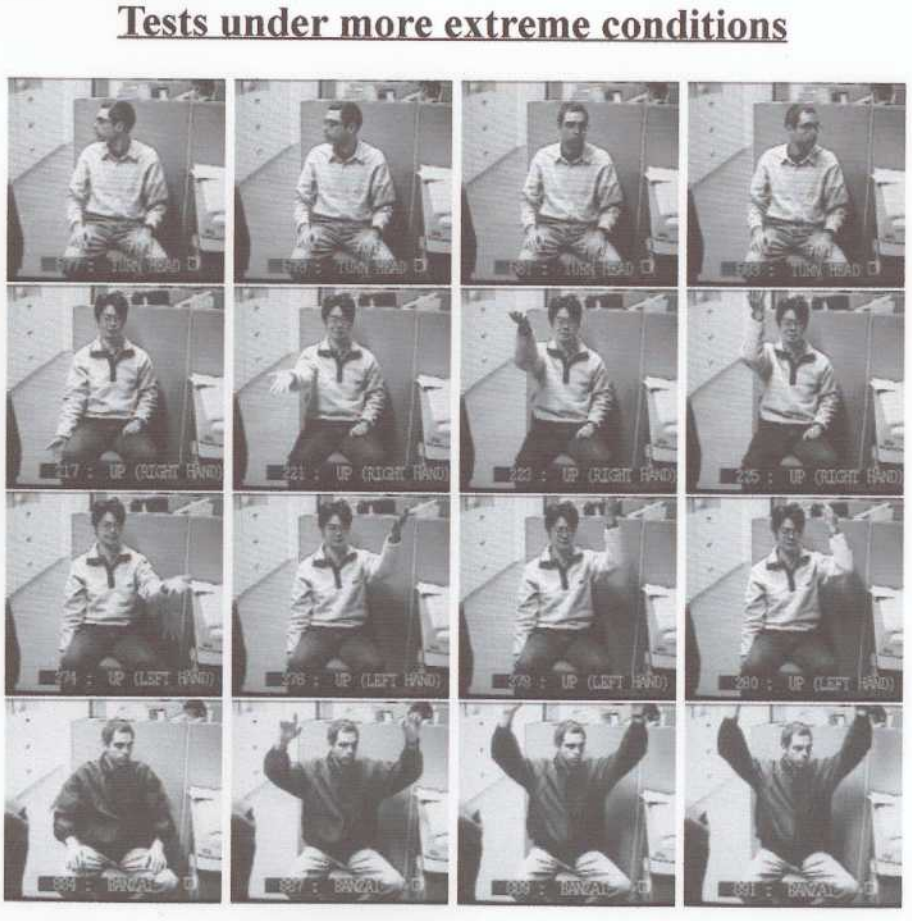

Figure 6-9. Snapshots of some of the gestures used in Experiment 6.2.5. Background and illumination conditions were chosen to be very different from those under which learning was performed (compare with Figs. 6-2 and 6-3) and video camera's angle and position were also significantly changed. Note also the shades formed as the subjects are performing the gestures.

Examples included:

(1) another person working with his computer in the background while the gestures were performed in the foreground;

(2) another person walking in the background;

(3) another subject occasionally moving near the subject performing the gestures;

(4) partial occlusion or truncation of body parts of the subject performing the gestures.

In all of the above-mentioned extreme cases, system's performance wasn't influenced significantly and average recognition rates were kept above 80%.