Stat Comput (2007) 17:57–70
DOI 10.1007/s11222-006-9004-9

Understanding the role of facial asymmetry in human face identification

Sinjini Mitra · Nicole A. Lazar · Yanxi Liu

Received: May 2005 / Accepted: September 2006 / Published online: 30 January 2007

Abstract Face recognition has important applications in forensics (criminal identification) and security (biometric authentication). The problem of face recognition has been extensively studied in the computer vision community, from a variety of perspectives. A relatively new development is the use of facial asymmetry in face recognition, and we present here the results of a statistical investigation of this biometric. We first show how facial asymmetry information can be used to perform three different face recognition tasks—human identification (in the presence of expression variations), classification of faces by expression, and classification of individuals according to sex. Initially, we use a simple classification method, and conduct a feature analysis which shows the particular facial regions that play the dominant role in achieving these three entirely different classification goals. We then pursue human identification under expression changes in greater depth, since this is the most important task from a practical point of view. Two different ways of improving the performance of the simple classifier are then discussed: (i) feature combinations and (ii) the use of resampling techniques (bagging and random subspaces). With these modifications, we succeed in obtaining near perfect classification results on a database of 55 individuals, a statistically significant improvement over the initial results as seen by hypothesis tests of proportions.

S. Mitra (✉)
Information Sciences Institute,University of Southern California,
4676, Admiralty Way, Suite 1001, Marina del Rey, CA 90292
e-mail: mitra@isi.edu

N. A. Lazar
Department of Statistics, University of Georgia

Y. Liu
The Robotics Institute, Carnegie Mellon University

Keywords Bagging · Classification · Facial asymmetry · Feature subset · Resampling

1 Introduction

Face recognition—a subclass of the broader problem of pattern recognition—is of increasing importance in recent years, due to its applicability in a wide variety of law enforcement and social arenas, such as matching surveillance photographs to mug shots, authentication checks at airports and ATMs, searching for missing children, and so forth. Since most of these applications are extremely sensitive in nature (for instance, catching criminals or terrorists), it is imperative to have highly accurate algorithms. Unlike in many applications where it is merely desirable to have a low rate of incorrect identification, face recognition requires it. The need for automatic algorithms is also evident, because it is difficult and time-consuming for a human to scan large facial databases (e.g. of missing children), especially in real time. As a result, automatic accurate face recognition has received much attention in the computer vision literature and numerous face identification algorithms have been developed.

All faces have a similar spatial layout and this makes face recognition a challenging task. Perhaps the greatest challenge is that images of a single individual may differ dramatically due to variations in orientation, color and illumination, or simply because the person's face looks different from day to day as a result of changes in make-up, facial hair, glasses, etc. On the other hand, researchers can take advantage of the fact that faces are rich in information about individual identity and mood; position relationships between parts of the face, including the eyes, nose, mouth and chin, as well as their shapes and sizes, are widely used as *features* for identification. One family of features that has only recently come into use in face recognition problems is *facial asymmetry*.

Facial asymmetry can arise in two ways—by external factors such as expression changes, viewing orientation and lighting direction, and by factors such as growth, injury and age-related changes. The latter is more interesting, since it is directly related to the individual face, whereas the former can be controlled to a large extent and may also be removed with the help of suitable normalization. Research has shown that the more asymmetric a face, the less attractive it is (Thornhill and Gangstad, 1999), but at the same time more recognizable, particularly in males (O'Toole, 1998). Human beings are so sensitive to naturally occurring facial asymmetry in recognizing individuals that a significant decrease in recognition performance has been observed when facial asymmetry is removed from images (Troje and Buelthoff, 1998). In fact, facial asymmetry has also been found to differ considerably between identical twins; Burke and Healy (1993), for example, reported significant differences in facial asymmetry parameters of monozygotic twins. This shows the potential of facial asymmetry as a useful biometric in practice.

The use of asymmetry in automatic human identification tasks started in computer vision with the work by Liu et al., (2002), which first showed that certain quantified facial asymmetry measures are indeed efficient in identifying people. This was followed by more extensive studies (Liu et al., 2003), and these early studies form the starting point for our current work, which presents a more rigorous study of facial asymmetry and its role in a variety of recognition problems. We begin with the role of asymmetry measures in human identification in the presence of extreme expression changes, followed by expression classification and male/female classification. We then explore the human identification problem in greater depth and propose the use of two different approaches for improving performance—combination of feature sets, and simple statistical resampling, which succeed in attaining very accurate identification results. Note here that, for the human identification task, we focus on the classification problem in this paper, where the goal is to identify a person who belongs to a particular database at hand. One potential application lies in law enforcement, where police or other agencies will be working with databases of known offenders, and one goal is to match suspect faces with targets from the database.

The paper is organized as follows. Section 2 provides a description of the data and Section 3 introduces the asymmetry features. The three different classification schemes are included in Section 4, along with a feature set analysis. Section 5 presents the classification results and Section 6 discusses the potential means of improving upon the human identification performance. We conclude with a discussion in Section 7.

2 Data

We use the same dataset as used by Liu et al. (2003), a part of the "Cohn-Kanade AU-coded Facial Expression Database" (Kanade et al., 1999). The Cohn-Kanade database features extreme expression changes captured under balanced lighting conditions, minimizing the external sources of asymmetry artifacts. As far as we know, this is the only database with these characteristics and hence is suitable for our analysis. The data are in the form of 165 video clips of 55 individuals displaying 3 emotions: joy, anger and disgust. Starting with a neutral expression, subjects gradually show an emotion, and so each clip exhibits the transition from a neutral face to one with the peaked form of one emotion. Every video clip is split into frames, each of which is a gray—scale image with pixels having numerical intensities ranging from 0 (black) to 255 (white). A total of 495 frames is used, 3 frames from each emotion for each subject: the most neutral (the first frame), the most peak (the final frame) and an intermediate expression (a middle frame). We use this smaller subset as our initial test-bed and hope to extend to a larger dataset in the near future.

The face images are first *normalized*. The goal of normalization is to align faces in such a manner that all the images are on a common plane, facilitating comparison. The normalization is based on an affine transformation and is explained briefly in Fig. 1. For more details on the exact procedure, the reader is referred to Liu et al. (2003). All the normalized images are of dimension 128 × 128 and they have a face

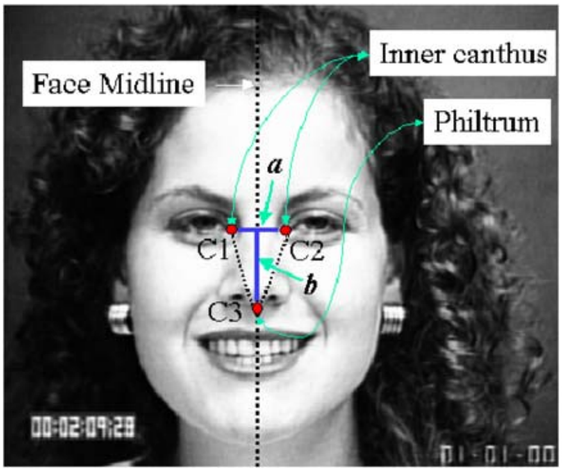


Fig. 1 Face image normalization—C_1 and C_2 denote the inner canthus of the two eyes, C_3 is the philtrum, a is the distance between C_1 and C_2 and b is the distance between the midpoint of $\overline{C_1C_2}$ and C_3. The normalization method determines the points C_1, C_2 and C_3 in each image for fixed distance values of a and b. [Courtesy Liu et al. (2003)]

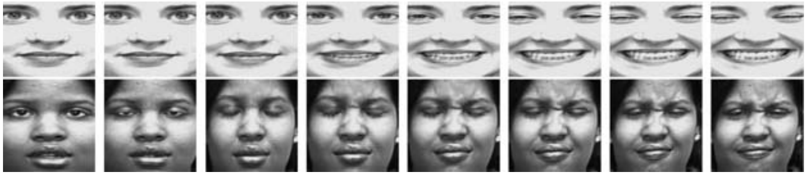


Fig. 2 Video sequences (8 normalized frames) of 2 subjects. The first subject (top row) is expressing joy and the second subject (bottom row), anger

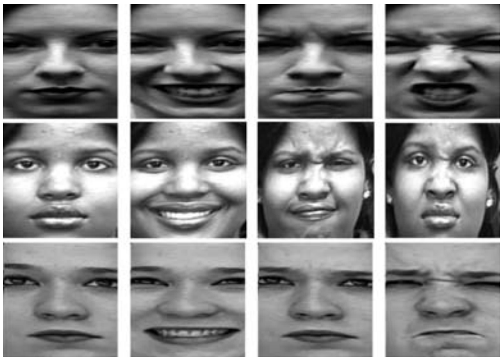


Fig. 3 Normalized expressions from 3 subjects. Each row represents one subject. Column 1 shows neutral expressions while columns 2–4 show joy, anger and disgust respectively

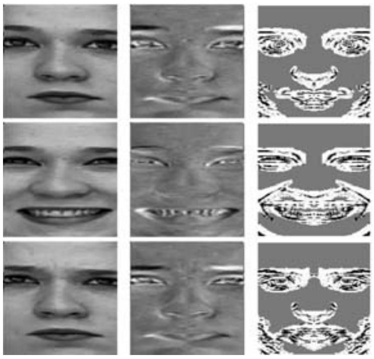


Fig. 4 The left column shows the normalized faces, the middle column shows the D-faces and the right column shows the S-faces. The three rows, from top to bottom, display neutral, joy and disgust expressions, respectively

midline so determined that each point on one side of the face has a corresponding point on the other (in other words, this is the "line of symmetry" in the face). Figure 2 shows sample video clips of two subjects expressing different emotions while Fig. 3 shows one frame for each of the three emotions and neutral expression for three subjects.

3 Facial asymmetry measurements

Following along the lines of Liu et al. (2003), we use two different representations of facial asymmetry in this paper. Using the face midline that was determined by the image normalization process, we define a coordinate system with the midline as the Y-axis and X-axis as the line perpendicular to it. If I denotes a normalized face and I' its vertiically reflected image along the midline, the two asymmetry measures are:

- **Density difference**: (*D-face*) Let $I(x, y)$ denote the intensity value of the image I at the coordinate location (x, y). Then

  $$D(x, y) = I(x, y) - I'(x, y),$$

  the intensity difference between the corresponding pixels from the two sides of the face, is the D-face value at that location. The higher the absolute value of $D(x, y)$, the more *asymmetrical* that point on the face is.
- **Edge orientation symmetry**: (*S-face*) Edges refer to significant and abrupt changes in the intensities of an image. An "edge" image I_e is obtained by detecting the edges in I, using a Laplacian-based method that is based on "zero-crossing" of the second derivative of the image (the point where it changes from positive to negative or vice versa is declared an edge). For more details, see the discussion in Lim, 1990. The advantage of this method over standard gradient-based ones (which only consider points whose gradients are above a fixed threshold) is that there is no check on the magnitude of the gradient and as a result, a larger number of points get declared as edge points. If I'_e denotes the vertically reflected image of I_e and ϕ the angle representing the difference in the orientations of the edges (that are detected automatically from the reflected images) at the two corresponding points on the face, the S-face value at the coordinate location (x, y) is

  $$S(x, y) = \cos\big(\phi_{I_e(x,y),I'_e(x,y)}\big).$$

  Since cosine is an even function, this measure is invariant to the relative orientation of the edges with respect to each other. The higher the value of $S(x, y)$, the more *symmetrical* that particular point on the face is.

Figure 4 shows the D-face and S-face for three different expressions of one person in our database. Both of these are of the same dimension as the normalized faces, that is, 128 × 128. However, the two sides of a D-face are exactly opposite of each other (same magnitude but different signs), whereas an S-face is the same on both sides (same magnitude and sign). Thus, half of each of these face contains all the relevant information, and it suffices to consider the half-faces alone. We construct three sets of features from each half-face, as follows:

- The values of each D and S-face are averaged over the 128 rows for each column in the half-face—this yields the *X-axis features*, known as D_{hx} and S_{hx} respectively. Each element of this feature vector corresponds to a horizontal line in the face going from the side of the face to the middle.
- The values of each D and S-face are averaged over the 64 columns for each row in the half-face—this yields the *Y-axis features*, known as D_{hy} and S_{hy} respectively. Each element of this feature vector corresponds to a vertical line on the face, from the top of the forehead to the bottom of the chin.
- Principal Component Analysis (PCA) is performed on the features of the half faces. Based on the eigenvalues, we keep the top 60 principal components for the D-face and the top 100 for the S-face since they explain 99% of the variation in their respective datasets. These feature sets are referred to as *D* and *S* respectively. Note that, PCA is used here as a dimensionality reduction technique only to determine the most discriminating combination of features from the feature sets as in standard statistical applications (Anderson, 1984).

Since each half D and S-face originally had $128 \times 64 = 8192$ features (same as the dimension of a half-face of a normalized image), the new feature sets are able to reduce the dimension of the problem, and at the same time summarize much of the essential information contained in the original 8192 pixel values.

4 Classification schemes

A commonly used human identification algorithm in computer vision is *Fisher Faces* (Belhumeur et al., 1997) which we will refer to as FF subsequently. This method uses a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to compute recognition features in a lower-dimensional subspace than the original feature space. Note that FF are not based on any kind of facial asymmetry information. For our problem, the top 25 Fisher Face features are computed from the normalized face images, and used as a benchmark for assessing the results based on our asymmetry measures. We chose 25 components since this proved to be optimal with respect to the trade-off between the number of components used and classification performance (that is, adding more components did not improve classification results significantly), as determined by cross-validation.

Table 1 Different feature sets and their dimensions

Features	D_{hx}	D_{hy}	S_{hx}	S_{hy}	D	S	FF
Number	64	128	64	128	60	100	25

For easy reference, we summarize the different feature sets used for this work, and the number of features contained in each in Table 1.

We are interested in three different identification tasks, namely,

1. Human identification in the presence of expression variations.
2. Classification of males/females.
3. Classification of the different expressions.

For **human identification**, we consider five experimental setups which offer different ways of representing training and test sets with widely varying expressions:

(i) Train on all frames from two emotions from all subjects and test on all frames from the third, that is, (a) train on anger and disgust and test on joy, (b) train on joy and disgust and test on anger, and (c) train on joy and anger and test on disgust.
(ii) Train on peak frames from the three emotions from all subjects and test on neutral ones, and vice versa.

These experimental setups help us assess the classification performance of our features on face images showing expressions that were not previously encountered. In real-life applications, a system will almost surely have to identify people with expressions it has not seen before, hence it is important to test the performance of the asymmetry features under such conditions. Here, all subjects are represented in the training and the testing samples. The goal is to correctly identify the subjects from their data in the test sample.

Identifying whether a person is a male or a female is important as that could potentially create smaller search domains leading to more efficient human identification. In other words, one could develop a male-specific classifier and female-specific classifier with the results from a sex classification routine which could then be used to identify a person of known sex—a task that we do not investigate in this paper but wish to pursue in the future. Our database has 15 males and 40 females. For **sex classification**, training is done on all 9 frames for a randomly selected subset of 8 males and 20 females and testing on all frames of the remaining 7 males and 20 females. Here, the goal is to classify each subject in the testing set as a man or a woman. The random selection of subjects for the training set is carried out 20 times (in order to remove selection bias), and final misclassification errors are obtained by averaging over the 20 iterations.

A person's expression is helpful in identifying his or her mood and mental state, and is often an individualized characteristic. People express emotions differently, which echoes human behavior and often helps in identification of a particular individual. Besides, facial expression recognition plays an important role in human computer interaction. For **expression classification**, training is done on the peak frames from the 3 emotions for a randomly selected subset of 30 people and testing on the peak frames from the 3 emotions of the remaining 25 people. Similar repetitions as in the sex classification case are performed for selecting the training sets and final errors computed in the same fashion. The middle frames are discarded for this purpose since some middle frames are closer to neutral expressions while others are closer to peak expressions, and this could potentially introduce bias in the classification results. Here, the goal is to classify the images in the test set according to the emotion being expressed.

The classifier that we use for all these identification tasks is Linear Discriminant Analysis (LDA; Anderson, 1984), along with a feature selection method known as Augmented Variance Ratio (AVR, Liu et al., 2002). Pattern recognition problems generally have a very large number of features—hundreds or even thousands are not unusual. The reason for this is that features are often generated automatically, without any knowledge of which ones are most likely to be meaningful or relevant. In our data, for instance, features have been generated from each pixel in a normalized face image. One task is then to discover which features are useful for classification. Moreover, massive datasets (images, microarrays, etc.) will often contain dependent features; the amount of independent information that is pertinent to the identification task is much less than that implied by the large number of features. Such features are completely redundant and they not only increase the complexity of the problem, but can also cause performance to deteriorate. For our face images, features correspond to different face parts which are spatially correlated and hence contain very small quantities of independent discriminating information. Ideally, those features which contribute to inter-class differences should have large variation between subjects and small variation within the same subject. AVR is a simple feature selection method and easy to compute. It compares within class and between class variances; at the same time it penalizes features whose class means are too close to one another. For a feature F with values S_F in a data set with C total classes, AVR is calculated as

$$AVR(S_F) = \frac{Var(S_F)}{\frac{1}{C}\sum_{k=1}^{C} \frac{Var_k(S_F)}{min_{j \neq k}(|mean_k(S_F) - mean_j(S_F)|)}},$$

where $mean_i(S_F)$ is the mean of the subset of values from feature F belonging to class i and provided $|mean_k(S_F) - mean_j(S_F)| \neq 0, \ \forall j, k$. When dealing with human identification, the individual subjects will form the classes, for expression classification the 3 expressions are the classes and finally, male and female form the 2 classes for the sex classification task.

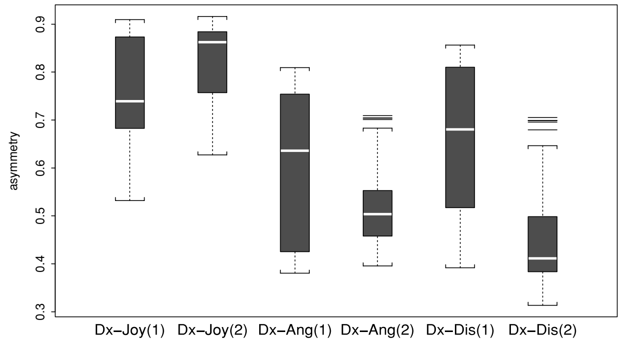


Fig. 5 Boxplots for the 3 emotions for 2 people—D_{hx}. 1 and 2 respectively denote the two people

Features are sorted in decreasing order of their AVR values and the one with the maximum value is first chosen. New features are then added according to a forward search algorithm if they improve the classification rate. A feature that worsens the classification or does not change it, given the ones that are already selected, is not included.

4.1 Feature analysis

In the feature analysis, we consider only the X-axis and Y-axis feature sets, and not the principal components. This is due to the fact that the latter are combinations of features and do not lend themselves to a natural physical interpretation.

We start with exploratory analysis in order to determine whether the asymmetry features based on the D and S-faces might be capable of providing efficient tools for the three identification goals stated above. Figures 5 and 6 show the boxplots of the feature values of all three emotions for two individuals in the database for datasets D_{hx} and S_{hy} respectively. These are constructed by averaging the values of each feature in a feature set over all the frames for each emotion separately. This pooling of all the features over frames is done only for the purpose of exploratory analysis and not for the subsequent classification routines. The plots indicate that there may exist some differences in the magnitude of asymmetry across people for the three expressions which we wish to exploit in identification tasks. The plots for the D_{hy} and S_{hx} are very similar and hence have not been included due to space considerations.

Figure 7 shows boxplots of the mean feature values for all the four feature sets, for men and women separately (mean over all frames from the people of the same sex). As can be seen in the figure, there exists some difference between the males and the females for all the four feature sets, males possessing an overall higher degree of facial asymmetry than

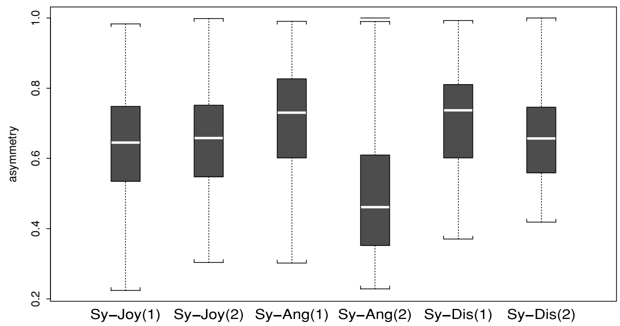


Fig. 6 Boxplots for the three emotions for two subjects—asymmetry measure S_{hy}. 1 and 2 denote the two individuals

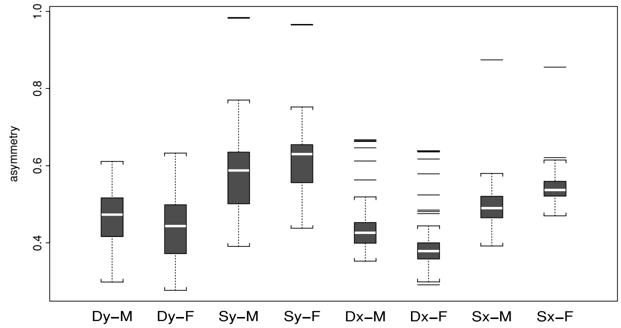


Fig. 7 Boxplots for the four asymmetry feature sets—D_{hx}, S_{hx}, D_{hy} and S_{hy}. M and F denote males and females, respectively

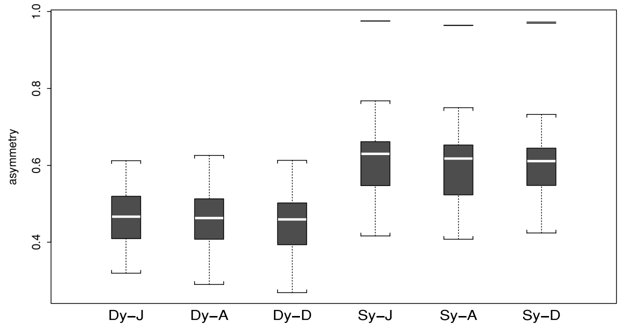


Fig. 8 Boxplots for the two asymmetry feature sets—D_{hy} and S_{hy}. J, A and D denote joy, anger and disgust, respectively

females (and, correspondingly low symmetry). This finding is consistent with results reported in Liu and Palmer (2003) on differences in the asymmetry measures between the sexes for three-dimensional human face images.

Figure 8 shows expression-wise boxplots of the mean feature values for the feature sets D_{hy} and S_{hy} (mean over all the frames showing the same emotion for all the people). Although the differences in facial asymmetry across the expressions do not seem to be as sharp as for the other two cases, it still seems worthwhile to investigate the role of asymmetry in expression classification.

Inspection of the boxplots in Figs. 7 and 8 reveals the presence of outliers in some of the feature sets (S-faces and D$_{hx}$ for sex). Almost all outliers are in the extreme right tail, representing larger than usual values of the respective features. Exploring the situation more carefully, we discover that the outlying S-face features correspond to the cheek region and such high symmetry occurs for most of the individuals in the dataset. In other words, almost all people have high symmetry around the cheeks, and indeed the symmetry in the cheeks is much higher than that in the other parts of the face. For D$_{hx}$, on the other hand, the outlying features were found to correspond to the forehead region which often possessed artificial asymmetry artifacts, some potential sources being edges arising from the cropping process and falling hair (see Fig. 9 for some sample images). Although edges appear in the images of many people (38 out of the 55), significant falling hair was found on only 6 people (subjects 10,15,16,23,28,33). Moreover, the D-face outliers were much smaller in magnitude than the S-face ones.

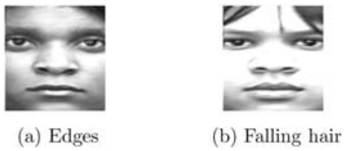
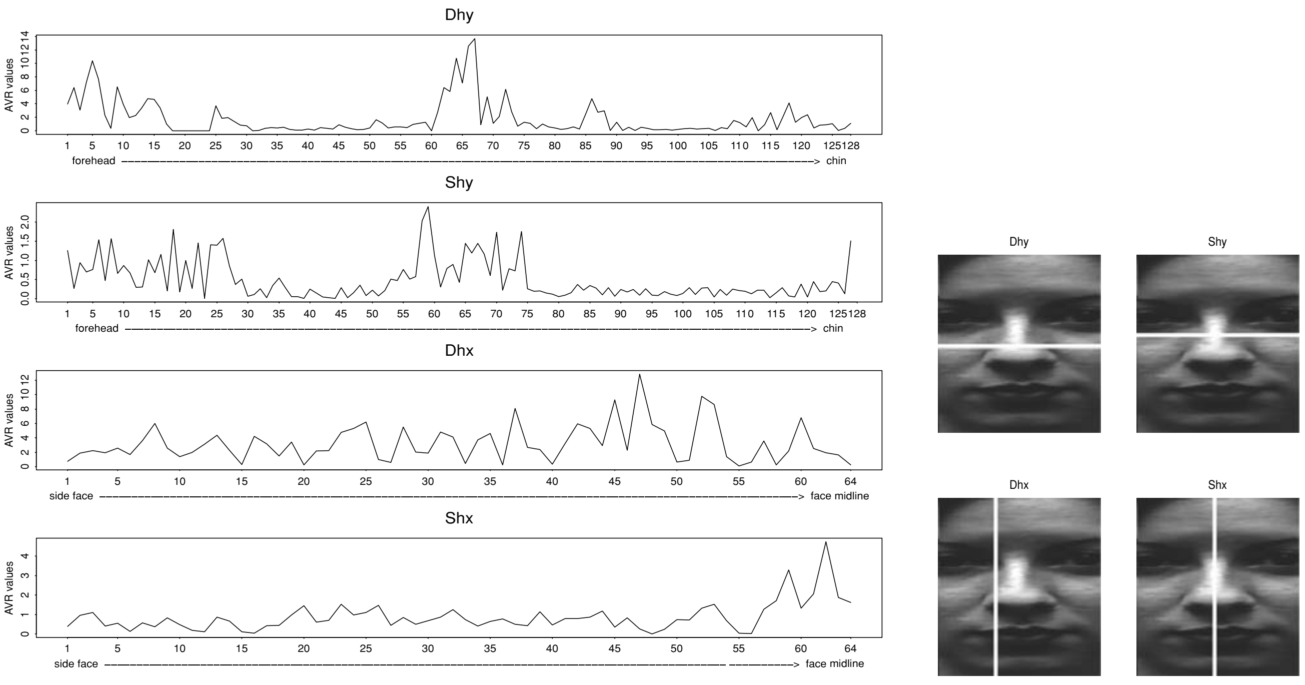

(a) Edges (b) Falling hair

Fig. 9 Sample images of people having artificial asymmetry artifacts in the forehead region

In the next step of the feature analysis we compute the AVR values of each of the feature sets for the three classification problems. The AVR values enable us to identify the particular parts of the face that help in the recognition process. Recall that the higher the AVR value of a feature (shown by the peak in the AVR plots), the more discriminating power it possesses and the more crucial it is for identification. Figure 10(a) shows AVR values of the four asymmetry feature sets for human identification. Features around the nose bridge and the forehead region play the most prominent role in human identification under expression variations. The most distinguishing feature from each set is indicated in Fig. 10(b).

The facial features that differ markedly in the amount of asymmetry between males and females are around the mouth, chin and above the eyes. The AVR plots for discriminating the two sexes are in Fig. 11(a); the features with the highest AVR values are marked in Fig. 11(b). Finally, Fig. 12(a) shows the AVR values of the features for expression classification, those with the highest AVR values being indicated in Fig. 12(b). A quick comparison with Fig. 10(b) reveals that unlike the case of human identification, which is greatly influenced by the bridge of the nose, the region around the mouth is discriminating across expressions. This indicates that asymmetry features corresponding to different facial parts contribute to these two apparently conflicting classification goals, which can henceforth be achieved with the help of the same set of features. This initial data analysis thus shows quite convincingly that facial asymmetry-based



(a) (b)

Fig. 10 (a) AVR values for the features from the four asymmetry feature sets for human identification. (b) The white line is drawn across the feature with the highest AVR value in each set, that is, the feature with the greatest discriminatory power

features have a potential for telling humans, expressions and the sexes apart.

5 Results

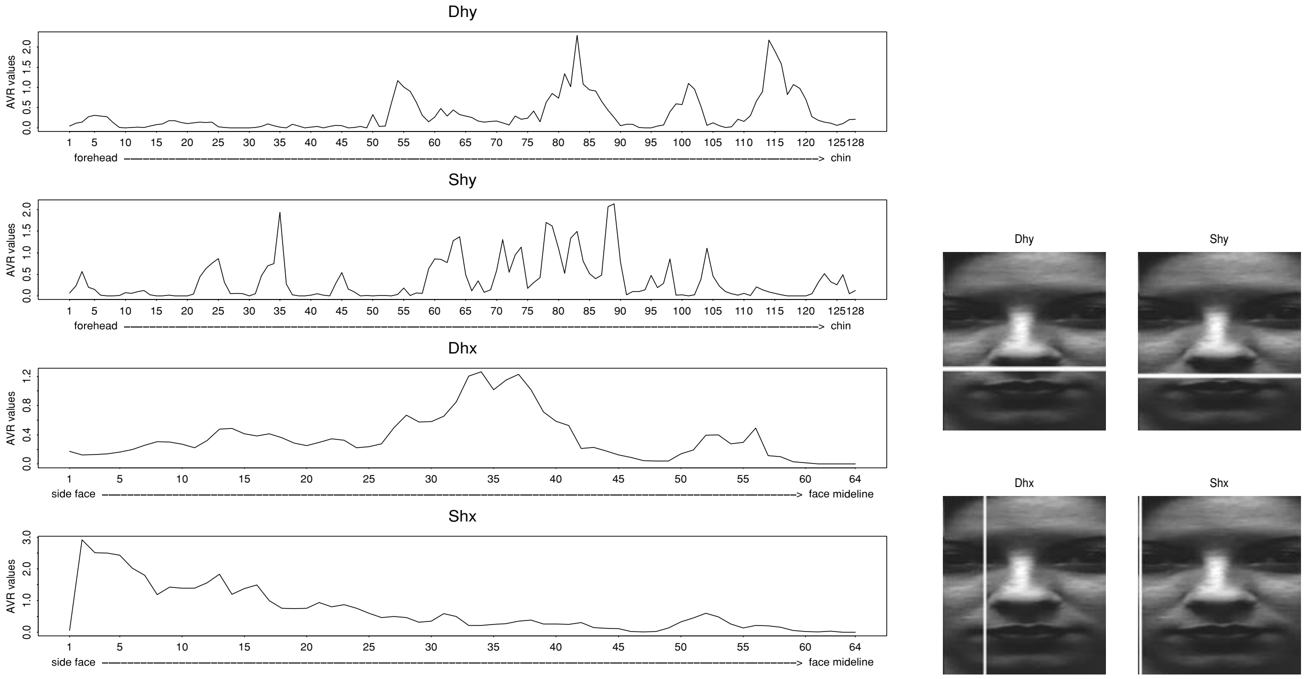
Motivated by the results of the exploratory analysis, we proceed to perform the three classification tasks using LDA as the classifier and AVR-based forward search as the feature selection criterion.

5.1 Human identification

The human classification results are shown in Table 2. Beside D, S and D_{hy}, none of the asymmetry feature sets produce satisfactory results. In particular, the X-axis features appear to be quite ineffective which demonstrates the fact that averaging over the rows of the face probably lost crucial asymmetry information by way of smoothing, whereas averaging over columns did not prove to be as harmful especially for D-face. The Fisher faces classifier, on the other hand, yields fair results which clearly shows room for improving upon our results, an issue that we investigate in the next section. The goal of any recognition technique is to achieve as near-perfect results as possible, and the next section presents ways of doing this by using features with greater discriminative power than either the asymmetry features or Fisher Faces alone. Among the five experiments, the neutral and peak frames are easiest to classify, yielding the lowest error rates. Of the emotions, the joy frames are hardest to classify. A possible explanation for this finding is that, when testing on joy, the training is done on the frames for anger and disgust, which have similar facial expressions (e.g. downturned mouth) and are different from joy (up-turned mouth). This is consistent with expectations since despite expression-invariance, the classification performance should deteriorate

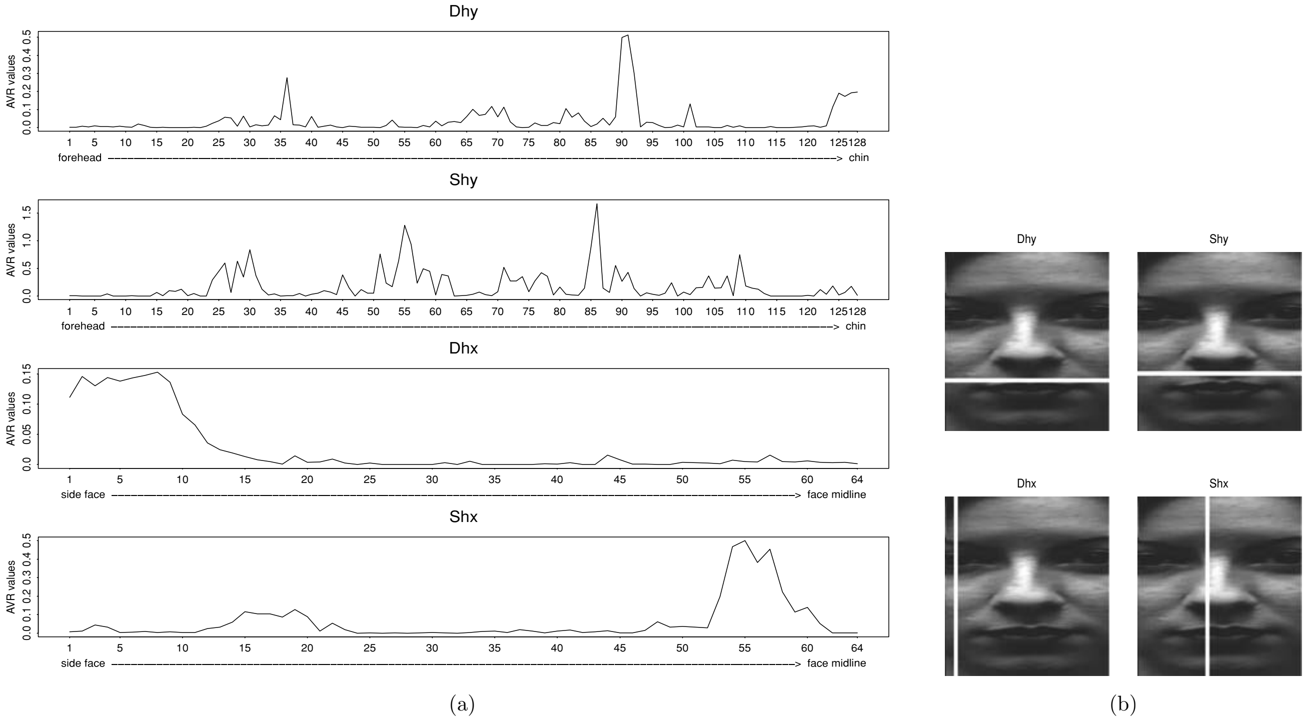
Table 2 Misclassification error rates for human classification, using the baseline method of LDA with AVR, computed as $\frac{M}{T} \times 100$, where M and T respectively are the number of cases misclassified and the total number of cases in each test set

Feature set/Test set	Joy (%)	Anger (%)	Disgust (%)	Neutral (%)	Peak (%)
D_{hy}	29.09	18.18	26.67	12.73	17.58
S_{hy}	42.42	36.97	43.03	25.45	29.70
D_{hx}	58.18	48.48	57.58	42.42	49.70
S_{hx}	65.45	60.00	61.82	48.48	53.94
D	18.18	12.72	10.30	3.03	3.03
S	21.82	24.24	18.79	4.85	10.91
FF	12.12	1.82	4.24	3.64	0.61



(a) (b)

Fig. 11 (a) AVR values for the different features for distinguishing men from women. (b) The lines are drawn at the facial features with the highest AVR values, that is, the features with the greatest ability to distinguish men from women



(a) (b)

Fig. 12 (a) AVR values for the different features for expression identification. (b) The lines are drawn at the facial features with the highest AVR values, that is, the features with the greatest ability to discriminate among the three expressions

Table 3 Misclassification error rates for sex classification (classwise and average)

Feature set	Male (%)	Female (%)	Average (%)	Std. error (%)
D_{hy}	35.74	20.56	24.70	6.76
S_{hy}	33.93	10.23	16.69	10.56
D_{hx}	46.80	34.60	37.93	5.43
S_{hx}	43.36	22.57	28.24	9.26
D	30.87	17.79	21.36	5.82
S	28.04	9.76	14.75	8.14
FF	23.48	6.53	11.15	7.55

Table 4 Misclassification error rates for sex classification using a "balanced" dataset with 15 males and 15 females (classwise and average)

Feature set	Male (%)	Female (%)	Average (%)	Std. error (%)
D_{hy}	13.56	15.51	14.53	2.46
S_{hy}	12.36	14.79	13.58	3.89
D_{hx}	30.56	32.90	31.73	5.87
S_{hx}	24.65	25.57	25.11	6.61
D	17.41	19.53	18.47	3.78
S	9.57	10.11	9.84	2.65
FF	4.68	7.89	6.28	3.65

as the tested expression deviates more and more from the trained expression, and the margin of this deterioration helps detect the robustness of the features under consideration.

5.2 Classification of sexes

Table 3 shows the sex classification results—classwise for males and females separately (columns 2–3), the average over the two classes (column 4) and the associated standard deviation over the two classes (column 5). Recall that the dataset is not balanced, containing 40 females and 15 males; the averages and standard deviations are computed in a weighted manner using the number of observations in each class as the corresponding weights (40 for class "female" and 15 for class "male"). The results show that, unlike human identification, S-face features are effective for distinguishing the sexes, although FF continues to outperform all other features. After FF, the lowest error rates are yielded by S_{hy} and S. Moreover, we also see that a female person has a lower chance of being misclassified as a male on an average, than that of a male being misclassified as a female. This may simply be an artifact of the composition of the sample, which would make the training more efficient for females than for males (since the classifier gets to see more examples of women). This is further corroborated by the fact that our exploratory analysis showed that male faces are more asymmetric and hence should be more recognizable and easier to classify than females. Moreover, Liu and Palmer (2003) observed that males are easier to classify than females on a dataset that was dominated by males. To explore this issue further, we generated balanced datasets in the following manner. We randomly selected a sample of 15 females out of the total 40, and performed classification of these along with the original sample of 15 males. The training was done with 8 males and 8 females and testing on the remaining 7 males and 7 females. In order to remove selection bias, the generation of 15 females was repeated 20 times and the final misclassification errors were obtained by averaging over these 20 iterations (Table 4). These results indicate that males have lower error rates than females as expected. Furthermore, the overall sex classification results improved significantly as a result of removing the bias created by the unbalanced sex ratio in our dataset. The standard error figures over the 20 repetitions measure the reliability of these error rates, and show that they do not fluctuate much with varying training samples.

5.3 Expression classification

Table 5 shows the expression classification results using the asymmetry-based features. We show the misclassification error rates emotion-wise (columns 2–4), the average error rate across all the three classes (column 5) and the associated standard deviations over the three classes (column 6). Unlike sex classification, here the standard deviation is computed in an unweighted manner as all the three expressions had the same number of observations. A comparison with the previous two problems shows that, unlike human identification and like sex classification, the S-face features are more efficient than the D-face features for distinguishing among the three emotions. Most of the other results, although significantly better than random guessing, are not that impressive, and FF again outperforms the rest. One thing worth noting here is that, for all feature sets except D_{hy}, the class "joy" has a lower error rate than the classes "anger" and "disgust". This may again be attributed to the fact that anger and disgust are expressed in a similar fashion (down-turned mouth, frown) and hence are more likely to be confused with each other than either is with joy, which has a contrasting expression style (up-turned mouth).

6 Improving human identification performance

The initial classification results outlined in the previous section are not very satisfactory in general, given the goal of attaining perfect to near perfect classification. There even seems to be room for improvement for the FF features, and in this section we explore means of improving upon them.

Table 5 Misclassification error rates for expression classification (classwise and average)

Feature set	Joy (%)	Anger (%)	Disgust (%)	Average (%)	Std. error (%)
D_{hy}	44.60	37.00	37.20	39.60	10.83
S_{hy}	14.00	20.80	21.20	18.67	7.25
D_{hx}	37.20	57.60	48.20	47.67	13.01
S_{hx}	39.20	51.80	41.40	44.13	11.62
D	25.80	42.80	41.60	36.73	13.32
S	15.60	17.80	20.00	17.80	7.93
FF	6.80	19.40	17.60	14.60	9.89

We focus on the human identification problem for now since this is the most important goal in practice, but the methods we describe can also be applied to any other classification problem. We consider two different approaches:

- Combining two or more feature sets together.
- Using statistical resampling techniques, namely bagging (Breiman, 1996) and RSM (Ho, 1998).

6.1 Combining feature sets

Table 6 shows misclassification error rates from combining some of the asymmetry feature sets together. The combination was performed simply by concatenating the features in the individual feature sets one after another in the order shown in the table. We indeed observe that adding more features improves the classification, as additional features supply more and different information, thus helping in identification. In fact, these results eventually get better than those obtained using FF alone. In some cases, even just combining two feature sets gives results that are at par with those from using FF (D+S, for instance). For several combinations, perfect classification is achieved for testing on the neutral and peak frames; if FFs are included along with the other feature sets, perfect classification is achieved for all test sets except for joy. These results vividly demonstrate how the error rates gradually drop when more features are combined, as can also be seen in Fig. 13. It is clear that asymmetry features can improve upon other classifiers by supplying complementary information that possesses greater discriminative power. In addition, simply combining asymmetry features results in a marked increase in performance in recognition tasks.

6.2 Resampling methods

Statistical resampling methods have been shown to be an effective and straightforward way of improving the performance of classifiers. Our goal in using resampling methods, however, goes beyond this. Aside from improving the classification results, we are also interested in the computational gain that can perhaps be achieved by obtaining high rates of correct classification while using a smaller number of features. To this end, we consider two different resampling methods—bagging and RSM—which are applied to the individual feature sets, and to some combinations.

6.2.1 Bagging

Bagging was introduced as a method for increasing the accuracy of *unstable* predictors, that is, predictors for which the results are significantly affected by small perturbations in the training set (Breiman, 1996). On the other hand, it is less effective if the underlying predictor is sufficiently stable, and can even do worse in such a scenario. According to Skurichina and Duin (1998, 2001), linear classifiers built on large training sets are stable. Hence, when the training sets are large, bagging will not improve results. Bagging is useless for very small training samples as well, since small training sets often represent the actual distribution poorly and the resultant classifiers are likely to be equally poor. However, when the training sample size is "critical" (the number of training samples is comparable to the number of features), linear classifiers can be quite unstable. So bag-

Table 6 Misclassification error rates from combinations of the asymmetry feature sets. (The FF results have been included for comparison purposes)

Feature set/Test set	Joy (%)	Anger (%)	Disgust (%)	Neutral (%)	Peak (%)
$D_{hy} + S_{hy}$	19.39	12.12	22.42	6.67	9.70
$D + S$	10.30	4.24	4.24	0.61	1.21
$D + S + S_{hy}$	8.48	1.82	4.24	0.61	1.21
$D + S + D_{hx} + S_{hx}$	6.67	1.21	3.64	0.61	0.00
$D_{hy} + S_{hy} + D_{hx} + S_{hx}$	17.58	10.91	18.18	5.45	7.27
$D{+}S + D_{hy} + S_{hy} + D_{hx}$	6.67	1.21	3.64	0.61	0.00
$D + S + D_{hy} + D_{hx} + S_{hx}$	7.88	1.21	3.64	0.00	0.00
$D + S + D_{hy} + S_{hy} + D_{hx} + S_{hx}$	6.67	1.21	3.03	0.00	0.00
FF	12.12	1.82	4.24	3.64	0.61
$FF + D + S + D_{hy} + S_{hy} + D_{hx} + S_{hx}$	2.42	0.00	0.00%	0.00	0.00

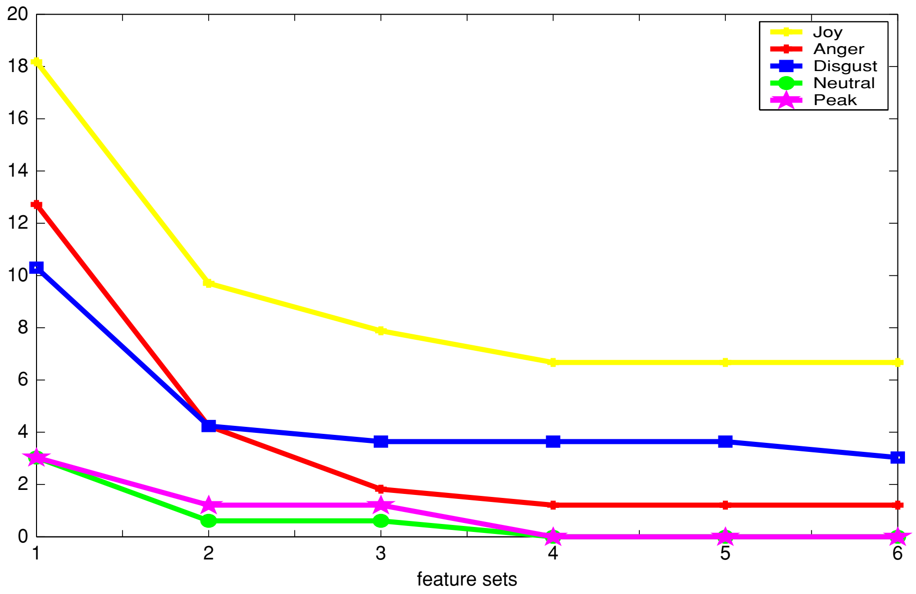


Fig. 13 Misclassification error rates for the different orders of combination of the asymmetry feature sets alone (D_{hy}, S_{hy}, D_{hx}, S_{hx}, D,S). The labels on the x-axis denote the number of feature sets in the respective combinations that produced the lowest error rates. For example, the lowest error rate for "joy", using only a single feature set, is 18.18%, attained with the D features; the lowest error rate for "joy" using two features sets is 10.3% (D+S); the lowest error rate for "joy" using three feature sets is 8.48% ($D + S + S_{hy}$), and so on

ging linear classifiers such as LDA might be beneficial for high-dimensional data in general.

The methodology of bagging consists of generating replications with replacement from the given training set and developing a classifier based on each of the samples by treating them as separate training sets. The final results are obtained by applying simple majority voting to the classification results from all samples. The number of replications is subjective and as Breiman points out, a greater number of classes usually calls for more resamples. We tried different numbers of replications from 10 to 100, the usual convention for statistical applications being 50. Since our problem is high dimensional, it seemed reasonable to believe that we will require a larger number of replications, and yet we decided to study this systematically. We report here only detailed results for D_{hy} due to space constraints but analogous phenomena are observed for all the other feature sets as well. These appear in Table 7 and a graphical version in Fig. 14. In general, different numbers of replications were found to be optimal (yielding the lowest error rates) for the different testing subsets even for the same feature set, ranging from 60–100. We did not observe significant improvement in the bagging results for replication sizes bigger than 100, while at the same time they increased the consumption of computational resources considerably (hence those results are not reported here). The standard errors associated with the error rates are also low and do not exhibit any noticeable pattern. The entire resampling procedure is repeated 20 times and final errors obtained by averaging over these 20 iterations.

The final bagging errors for all the feature sets and some of their combinations are summarized in Table 8, which shows an appreciable improvement in the classification performance. We did not apply bagging on test sets for which the baseline results were perfect (0% error rate), since no further improvement is possible.

6.2.2 *Random Subspace Method (RSM)*

The Random Subspace Method (RSM), introduced by Ho (1998) samples from the feature space. If there are originally p features, one randomly selects $p^* < p$ and a classifier is built on the p^*-dimensional space (subspace of the original p-dimensional space) using all the training samples. This process is repeated and a majority voting technique yields the

Table 7 Bagging misclassification rates and corresponding standard deviations over 20 repetitions (%) for D_{hy}. The figures in bold represent the optimal results among all the replication sizes

B Original	Test on joy 29.09%	Test on anger 18.18%	Test on disgust 26.67%	Test on neutral 12.73%	Test on peak 17.58%
10	20.33 ± 0.8	14.30 ± 0.6	22.12 ± 0.6	6.39 ± 0.2	11.64 ± 0.3
20	17.64 ± 0.5	10.73 ± 0.3	20.36 ± 0.4	4.39 ± 0.2	8.76 ± 0.2
30	17.06 ± 0.3	10.15 ± 0.3	20.09 ± 0.3	3.91 ± 0.1	8.58 ± 0.2
40	16.97 ± 0.3	9.15 ± 0.2	20.64 ± 0.3	3.15 ± 0.1	8.48 ± 0.2
50	16.61 ± 0.3	10.78 ± 0.1	20.61 ± 0.5	3.39 ± 0.1	7.88 ± 0.1
60	16.76 ± 0.3	**8.88 ± 0.1**	**19.61 ± 0.2**	3.15 ± 0.1	7.94 ± 0.1
70	16.45 ± 0.1	10.75 ± 0.2	20.85 ± 0.5	3.36 ± 0.1	**7.61 ± 0.6**
100	**16.30 ± 0.6**	10.76 ± 0.1	20.03 ± 0.2	**2.91 ± 0.2**	7.97 ± 0.2

Table 8 Misclassification error rates and standard deviations (%) from applying bagging to the asymmetry feature sets. The standard errors have been rounded to the nearest one decimal place

Features/Test set	Joy	Anger	Disgust	Neutral	Peak
D_{hy}	16.30 ± 0.6	8.88 ± 0.1	19.61 ± 0.2	2.91 ± 0.2	7.61 ± 0.6
S_{hy}	23.45 ± 1.1	24.18 ± 1.0	24.64 ± 0.4	10.91 ± 0.2	14.42 ± 0.7
D_{hx}	52.06 ± 1.8	41.18 ± 1.8	49.39 ± 1.8	31.91 ± 1.8	33.03 ± 1.8
S_{hx}	55.15 ± 1.3	44.94 ± 1.3	49.01 ± 1.4	32.85 ± 1.1	39.64 ± 0.9
D	11.21 ± 0.7	5.15 ± 0.7	5.58 ± 0.3	0.79 ± 0.1	1.58 ± 0.2
S	13.33 ± 0.3	11.73 ± 1.0	10.27 ± 0.4	2.30 ± 0.1	4.58 ± 0.3
FF + D	3.18 ± 0.5	0.64 ± 0.0	0.61 ± 0.1	–	–
$D_{hy} + S_{hy} + D_{hx} + S_{hx}$	8.36 ± 0.5	3.00 ± 0.3	8.51 ± 0.3	1.21 ± 0.1	0.54 ± 0.1
$D + S + D_{hy} + S_{hy} + D_{hx} + S_{hx}$	5.06 ± 1.1	1.12 ± 0.2	0.76 ± 0.1	–	–
$FF + D + S + D_{hy} + S_{hy} + D_{hx} + S_{hx}$	0.24 ± 0.2	–	–	–	–
FF (no bagging)	12.12%	1.82%	4.24%	3.64%	0.61%

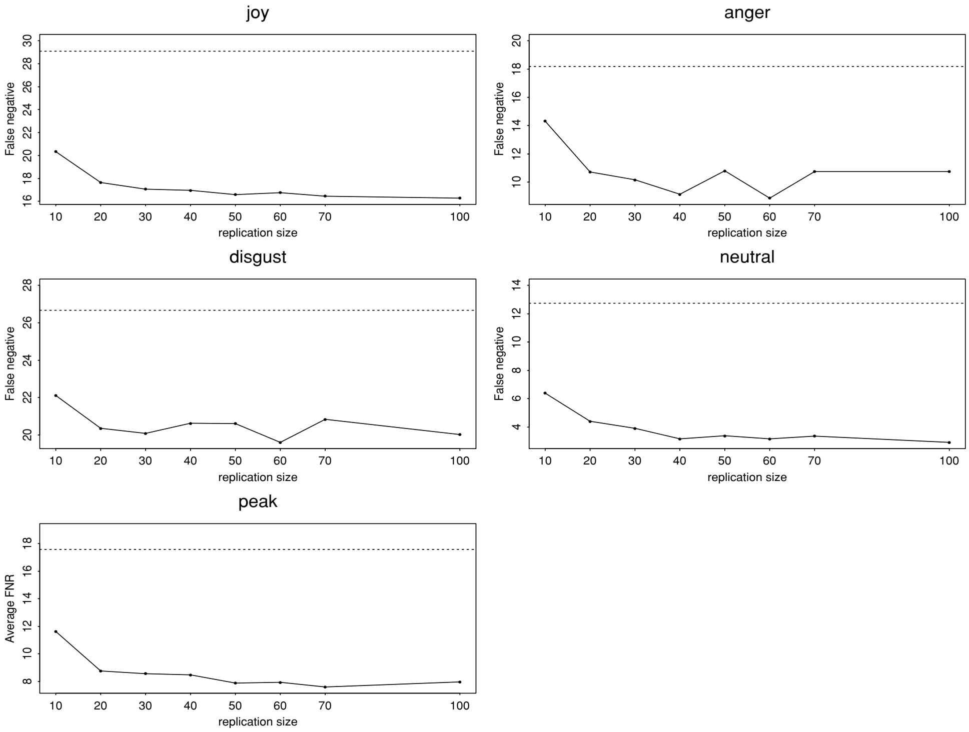


Fig. 14 The bagging misclassification errors for the five testing subsets as compared to the original one for the D_{hy} dataset. The dotted line represents the original error in each case

final classifier. The subspace dimensionality is thus smaller than that of the original feature space, but the number of training objects remains the same. RSM is known to perform well when there is some redundancy present in the feature space, otherwise, it is not guaranteed to give better than baseline results (Skurichina and Duin, 2001). This is because redundancy in the feature space is likely to deteriorate performance and RSM helps remove this redundancy by way of sampling repeatedly. Thus RSM performs poorly when all features are informative and there is no significant redundancy.

Table 9 shows the error rates obtained as a result of applying RSM to the asymmetry features. In each case, p^* is 50% of p, the total number of features available (a common convention). As in bagging, we repeat the entire procedure 20 times and the final errors are obtained by averaging over these 20 repetitions. D_{hy}, S_{hy}, D_{hx}, S_{hx} (and their combinations) have considerable redundancy and we find that RSM improves over the baseline LDA results, as expected. By way of contrast, RSM often deteriorates in performance when some feature extraction has been performed, a serious restriction on the applicability of this resampling method. The D and S principal component feature sets are obtained by a dimension-reduction technique and are orthogonal to each other; they have no redundancy and it is therefore not expected that RSM would lead to better results in those cases. Indeed, we found that there is either a deterioration or only a

Table 9 Misclassification error rates and standard errors (%) from applying RSM to the asymmetry feature sets using only 50% of the features in each feature set. The standard errors have been rounded to the nearest one decimal place

	Joy	Anger	Disgust	Neutral	Peak
D_{hy}	19.94 ± 0.7	11.67 ± 0.7	23.09 ± 0.4	4.79 ± 0.3	11.58 ± 0.2
S_{hy}	29.57 ± 1.0	28.54 ± 1.0	31.51 ± 0.9	15.51 ± 0.2	22.12 ± 0.9
D_{hx}	54.64 ± 1.2	46.97 ± 1.1	52.58 ± 1.6	35.61 ± 0.8	46.30 ± 1.3
S_{hx}	63.18 ± 0.2	54.45 ± 0.3	55.42 ± 0.5	38.58 ± 0.2	39.64 ± 0.1
$\mathrm{D}_{hy} + \mathrm{S}_{hy} + \mathrm{D}_{hx} + \mathrm{S}_{hx}$	10.56 ± 1.3	5.15 ± 0.2	10.42 ± 0.8	1.21 ± 0.1	1.85 ± 0.1
FF (no bagging)	12.12%	1.82%	4.24%	3.64%	0.61%

marginal improvement (which might be due to chance) over the original results for D and S (and also in the combinations with them), hence we do not report these results here.

6.2.3 Comparison of resampling results

Both the resampling methods produce good results with fewer feature sets, the prime motivation for applying them in the first place. For example, applying bagging and RSM to the combination of the four asymmetry feature sets, D_{hy}, S_{hy}, D_{hx}, S_{hx}, we are able to achieve considerably lower error rates than the original FF for testing on three out of the five subsets (joy, neutral, peak); these improvements are highly statistically significant, with p-values close to zero. Furthermore, when combining asymmetry features with FF, we find that adding just the D-face principal components produces very impressive results that are not (statistically) significantly different from combining with all the six asymmetry features. This is a definite improvement considering that the feature sets are high-dimensional; the smaller the number of features one needs to use the better. Moreover, the standard errors of the misclassification rates for both the resampling methods (computed over the 20 repetitions in each case) are quite low, and this establishes the reliability of these estimates.

We also note that RSM does not achieve improvement over the original results as uniformly as bagging. Only for the S_{hy} feature set and the combination of the four asymmetry face feature sets, do we observe statistically significant improvements for all the five testing subsets. Such significant improvements for the other datasets are limited (p-values close to 1). Moreover, we could not apply RSM to all of our feature sets. The relative lack of flexibility of RSM, compared to bagging, brings its efficacy in this particular application into question.

7 Discussion

The baseline method of LDA yielded results for this 55-class problem that were not entirely satisfactory, given our high standard of perfect, or near perfect, classification, although the initial exploratory analysis indicated that asymmetry measures have the potential for recognizing people under expression variations, as well as for expression and sex classification tasks. The asymmetry features are simple to compute, a fact that enhances their practical utility. On the other hand, computation of these features requires high resolution facial images with extrinsic sources of variation (pose and lighting) carefully controlled, which may limit their use in real-time surveillance applications.

We have also shown that classification performance of these asymmetry measures can be improved by using two very simple techniques—combinations of feature sets and statistical resampling. Both these methods have proved successful in producing good results that are as good as or better than the Fisher Face features. The resampling methods not only improve upon the baseline LDA results, but do so with a small number of feature sets. Both these aspects make the resampling methods quite attractive to users from the standpoint of accuracy as well as efficiency. These are extremely important issues given the sensitive nature of most face recognition applications, where misclassification can have a drastic impact and it is imperative to have very accurate algorithms. Furthermore, the ease of their implementation also makes real-time application a possibility. The algorithms are quick to execute, and require very little in terms of additional resources or time, thus increasing their scope of application. Moreover, such good results also bring out the true potential of the asymmetry features in expression-invariant human recognition. Inspired by these results, our immediate future direction is to apply these improvement methods to both expression and sex classification.

It is also worth recalling here another well-known resampling technique known as boosting (Freund and Schapire, 1997), which has been used successfully in a variety of pattern recognition problems (Viola and Jones, 2001). According to Skurichina and Duin (2000), boosting does depend on the instability of the classifier and is not beneficial for linear classifiers such as LDA. Hence, we did not attempt to use it for our experiments, since we wished to use the same base classifier for all experiments in order to have a fair comparison. One classifier for which boosting can be beneficial is the Nearest Mean Classifier, which we plan to investigate in the future.

Note that we use the same subset of the Cohn-Kanade database that was used by Liu et al. (2003) as our initial testbed, since our analyses were intended to provide a firmer basis and evaluation for this earlier work, and hence use their results for fair comparison. This data subset had 55 individuals and it would definitely be beneficial to apply our methods to a larger database with more people; we wish to pursue this more fully in the future. This would further strengthen the robustness of our approach although even for the smaller dataset, all our classification results (human, sex and expression) were observed to be significantly better than those obtained from mere random guessing in a statistical sense. We were also able to achieve better results than Fisher Faces when using feature combination and resampling techniques.

At this point, a reader might wonder whether our results will also be valid for spontaneously-produced expressions. The people in this study produced the emotions on demand—that is, when asked they started with a neutral expression which gradually evolved into a joyous or angry or disgusted expression. However, this does not happen in practice; surveillance cameras usually capture a face with a spontaneous expression. Indeed, according to Hager and Ekman (1985), posed facial expressions are more asymmetric than spontaneous ones, and so our results might be providing a biased estimate of the extent to which facial asymmetry can truly aid in face recognition. It would thus be useful to investigate how classification results change when using genuine expressions.

At the end, we wish to mention briefly the scope of generalizing our results. We are interested in the asymmetry caused by the actual facial structure, which depends on growth-related factors and hence it may be interesting to study whether it is more difficult to identify younger people or older people based on their facial asymmetry. Similarly, it may be useful to determine whether people of certain ethnic origins are easier to identify based on facial asymmetry than others. Although we expect our techniques to yield fairly good results on databases with images of people with diversified demographics, we intend to explore this issue in greater depth by using a larger database.

Another direction of research that we intend to pursue in the future is exploring the efficacy of the asymmetry features in verification. So far we have been concerned with training and testing the face images of the same group of individuals but an interesting scenario will be to test if our algorithms are able to detect a test face for which corresponding training samples are not available. This will have an important application in identifying the people that are on the do-not-fly list at airports as a measure of security check. This is expected to boost the utility of these measures in practice considerably.

Acknowledgment The authors wish to thank Jeff Cohn and Karen Schmidt for providing the database, and the Associate Editor and the reviewers for various helpful comments and suggestions.

References

Anderson T.W. 1984. An Introduction to Multivariate Statistical Analysis, 2nd Edn. Wiley, New York.

Belhumeur P.N., Hespanha J.P., and Kriegman D. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transaction on Pattern Analysis and Machine Intelligence 19(7): 711–720.

Breiman L. 1996. Bagging predictors. Machine Learning 24(2): 123–140.

Burke P.H. and Healy M.J. 1993. A serial study of normal facial asymmetry in monozygotic twins. Annals of Human Biology 20(6): 527–534.

Freund Y. and Schapire R. 1997. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55(1): 119–139.

Hager J. and Ekman P. 1985. The asymmetry of facial actions is inconsistent with models of hemispheric specialization. Psychophysiology 22: 307–318.

Ho T.K. 1998. The random subspace method for constructing decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8): 832–844.

Kanade T., Cohn J.F., and Tian Y.L. 1999. Comprehensive database for facial expression analysis. In: 4th IEEE International Conference on Automatic and Gesture Recognition. Grenoble, Fr.

Lim J.S. 1990. Two-Dimensional Signal and Image Processing. Prentice Hall, New Jersey.

Liu Y. and Palmer J. 2003. A quantified study of facial asymmetry in 3d faces. In: Proceedings of the 2003 IEEE International Workshop on Analysis and Modeling of Faces and Gestures.

Liu Y., Schmidt K., Cohn J., and Mitra S. 2003. Facial asymmetry quantification for expression-invariant human identification. Computer Vision and Image Understanding Journal 91(1/2): 138–159.

Liu Y., Schmidt K., Cohn J., and Weaver R.L. 2002. Human facial asymmetry for expression-invariant facial identification. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FG'02).

O'Toole. 1998. The perception of face gender: the role of stimulus structure in recognition and classification. Memory and Cognition 26(1): 146–160.

Skurichina M. and Duin R.P.W. 1998. Bagging for linear classifiers. Pattern Recognition 31(7): 909–930.

Skurichina M. and Duin R.P.W. 2000. Boosting in linear discriminant analysis. In: Lecture Notes in Computer Science. Vol. 1857. Springer-Verlag, Berlin.

Skurichina M. and Duin R.P.W. 2001. Bagging and the random subspace method for redundant feature spaces. In: Lecture Notes in Computer Science, Vol. 2096. Springer-Verlag, Berlin.

Thornhill R. and Gangstad S.W. (1999) Facial attractiveness. Transactions in Cognitive Sciences 3(12): 452–460.

Troje N.F. and Buelthoff H.H. 1998. How is bilateral symmetry of human faces used for recognition of novel views? Vision Research 38(1): 79–89.

Viola P. and Jones M. 2001. Robust real-time object detection. In: International Conference of Computer Vision.