# Classical and Resampling Methods for Face Recognition based on Quantified Asymmetry Measures

Sinjini Mitra, Nicole A. Lazar, Yanxi Liu

**Abstract**

Face recognition has important applications in psychology and biometric-based authentication, which increases the need for developing automatic face identification systems. Psychologists have long been studying the link between symmetry and attractiveness of the human face, but based on qualitative human judgment alone. The use of objective facial asymmetry information in automatic face recognition tasks is relatively new. The current paper presents a statistical analysis of the role of facial asymmetry measures in face recognition, under expression variation. We first describe a baseline classification method and show that the results are comparable with those based on certain popular (non-asymmetry based) classes of features used in computer vision. We find that facial asymmetry further improves upon the classification performance of these popular features by providing complementary information. Next, we consider two resampling methods to improve upon the baseline method used in previous work, and present a detailed comparison study. We demonstrate that resampling methods succeed in obtaining near perfect classification results on a database of 55 individuals, a statistically significant improvement over the baseline method. Results regarding the role of asymmetry of different parts of the face in distinguishing between individuals, expressions and between males and females are also reported as additional aspects of the study.

**Keywords:** asymmetry, bagging, classification, expressions, feature subset, random subspace, 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 (police officers are reluctant to falsely accuse an innocent person of a crime, investigators do not want to plant false hope in the minds of the parents of missing children), it is imperative to have highly accurate algorithms. In other words, unlike in some applications where it is merely desirable to have a low rate of incorrect identification, face recognition requires it. As a result, automatic accurate face recognition has received much attention in the computer vision literature.

An obvious tactic in this quest for precise classification schemes is to use characteristics of the face itself in the identification process. Position relationships between parts of the face (in particular, eyes, nose, mouth and chin), as well as their shapes and sizes, are highly individualized, a fact that has been exploited by defining measures or "features" which capture these essential aspects of physiology. One family of features that has only recently come into use in face recognition problems is *facial asymmetry*.

There are two kinds of facial asymmetry - intrinsic and extrinsic. The former is caused by growth, injury and age-related changes, while the latter is affected by viewing orientation and lighting direction. Intrinsic asymmetry is the more interesting of the two, since it is directly related to the individual face whereas extrinsic asymmetry can be controlled to a large extent. The more (intrinsically) asymmetric a face, the less attractive it is (Thornhill and Gangstad, 1999), but at the same time more recognizable (O'Toole, 1998; Troje and Buelthoff, 1998). In fact, we 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. These findings suggest that measures of facial asymmetry might enhance the performance of automatic face recognition routines. Research in this area started in the computer vision community in 2001 with the seminal work by Liu (Liu et al, 2002), which first showed that certain facial asymmetry measures are efficient in identifying people in the presence of expression variations. This was followed by more extensive studies (Liu et al, 2003), which investigated further the role of different types and locations of asymmetry features both on their own, and as a complement to other popular classes of features.

These studies lay the ground for our current work, which shows how some commonly used statistical resampling methods such as bagging (Breiman, 1996) can help achieve appreciably better classification results over the baseline method used in previous research, with relatively small extra resources. The enhanced methods we present in this paper attain near perfect classification, although the number of classes is large (55 individuals to identify), much larger than is usual in statistical applications. The marriage of methods from the two disciplines, computer vision and statistics, surpasses what either is able to achieve on its own.

The outline of the rest of the paper is as follows: Section 2 describes the baseline methodology and the resampling techniques (bagging and random subspace) used in the improved classification procedure. Section 3 provides a description of the data and Section 4 lays down the experimental setup. Section 5 contains a summary of the classification results using the baseline method, and the results from the two resampling methods. Section 6 presents some analyses of the selected feature subsets. A general discussion appears in Section 7.

# 2 Methods

In section 2.1 we introduce the classifier which was the primary classifier for the work reported in Liu et al (2003) - Linear Discriminant Analysis (LDA; Anderson, 1984) with Augmented Variance Ratio (AVR; Liu et al, 2002). Section 2.2 then describes the two re-sampling methods that we consider for the present work. Both resampling methods use LDA with AVR as the base classifier.

## 2.1 The Baseline Method - LDA with AVR

The initial analysis consists of using LDA as the classification method and AVR as the feature selection criterion. Computer vision problems generally have a very large number of features - hundreds or even thousands are not unusual. This is because, for many problems, features are generated automatically without any knowledge of which ones are most useful. One task is then to discover which of these features are useful for purposes of classification. The primary reason for performing feature selection is therefore to reduce the amount of redundancy in the dataset. The idea behind using variance ratios is that those features which contribute to inter-class differences should have large variation between subjects and small variation within the same subject. Thus, the ratio of between-class variance to within-class variance should be large. The AVR compares within class and between class variances and 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 $F$ which belong to class $i$. The augmented variance ratio thus imposes a penalty on features which may have small intra-class variance but which have close inter-class mean values. For our problem, the classes refer to the subjects.

Features are sorted in decreasing order of their AVR values and selected according to a forward search algorithm. We start with the feature having the maximum value of AVR and then add on new ones if they improve the classification. A feature that worsens the classification or does not change it, given the ones that are already selected, is not included.

## 2.2 Resampling Methods

We consider two resampling methods, with the objective of improving upon the initial classification results: (1) Bagging and (2) Random Subspace Method (RSM).

### 2.2.1 Bagging

Bagging was introduced by Breiman as a method of increasing the accuracy of certain predictors. It is a fairly simple way to improve upon an already existing method (Breiman, 1996). Bagging is mostly beneficial when the underlying predictor is unstable, that is, if small perturbations of the learning set cause significant changes. On the other hand, it is less effective if the underlying predictor is sufficiently stable, and can even do worse in such a scenario.

The methodology of bagging consists of bootstrapping training samples repeatedly and developing a classifier based on each of the bootstrap samples by treating them as separate training sets. The final results are obtained by combining the classification results from all samples using a simple majority voting technique, which involves assigning each observation to the class to which it is classified in the maximum number of instances among all the bootstrap samples constructed.

Breiman (1996) notes that the success of bagging depends on several factors, including the training sample size, the choice of the base classifier and the potential ability of the chosen base classifier to solve the problem. The number of replications, however, is subjective and depends on the available data, the problem at hand and also the available resources. In general, more replicates are required with an increasing number of classes in a classification problem.

Skurichina and Duin (1998, 2001) discuss several perspectives on applying bagging to linear classifiers such as LDA. According to these authors, linear classifiers built on large training sets are stable, and hence bagging LDA will tend to not be useful. Bagging is useless for very small training samples as well, since small training sets often represent the distribution of the entire dataset poorly; classifiers are likely to have correspondingly poor performance. However, when the training sample size is "critical", that is, when the number of training objects is comparable with the number of features, linear classifiers such as LDA can be quite unstable. Thus, bagging LDA might be beneficial for data that are high dimensional, in general. Some of our facial asymmetry datasets are quite high dimensional and hence bagging should be applicable.

### 2.2.2 Random Subspace Method (RSM)

The Random Subspace Method (RSM) was introduced by Ho (1998). This method bootstraps features from the feature space, instead of from the training samples. In other words, if we have a data matrix with features along the columns and samples along the rows, then bagging samples from the row dimension whereas RSM samples from the column dimension.

If there are originally $p$ features, one randomly selects $p^* < p$, thus obtaining a $p^*$-dimensional random subspace of the original $p$-dimensional feature space. The classifier is then built on each of these $p^*$-dimensional spaces and an aggregating technique such as simple majority voting is used to get the final classifier. The subspace dimensionality is thus smaller than that of the original feature space, but the number of training objects remains the same. When the data have many redundant features, one may obtain better classifiers in random subspaces than in the original feature space. The combined decision of such classifiers may be superior to that of a single classifier constructed on the original training set in the complete feature space. Hence RSM does not perform well when all features are informative and there is no significant amount of redundancy. By design, facial asymmetry features have much redundancy and hence RSM has the potenial to be effective.

# 3 Data

The dataset that we consider here is the same one as used in Liu et al (2003), the "Cohn-Kanade AU-coded Facial Expression Database" (Kanade, et al, 1999). It consists of 165 video clips displaying three emotions - joy, anger and disgust - of 55 subjects. Starting with a neutral expression, subjects gradually show joy, anger or disgust. Each clip therefore exhibits a transition from a neutral face to one with the peak form of a particular emotion. Each video clip is broken down into frames consisting of a $640 \times 480$ pixelized gray-scale image. Pixels in the frame all have numerical intensities ranging from 0 (black) to 255 (white). We use a total of 495 frames, three frames from each of the three emotions for each subject. The three frames from each emotion sequence are the most neutral (the beginning one), the most peak (the final one) and the middle (the one in the middle of the entire sequence).

## 3.1 Face Image Normalization

The goal of normalization is to establish a common co-ordinate system to facilitate comparison of the different faces. A *face triangle* is defined as the triangle formed by connecting three points: the inner canthus of each eye, denoted ($C_1$,$C_2$), and the philtrum, denoted $C_3$ (Figure 1). Let $\overline{C_1C_2}$ denote the line segment joining the points $C_1$ and $C_2$. Each image is then *normalized* using the following steps:

(i) **rotation**: rotate $\overline{C_1C_2}$ into a horizontal line segment

(ii) **X-scaling**: scale $\overline{C_1C_2}$ into length $a$

(iii) **X-skewing**: skew the face image horizontally so that $C_3$ is located on the perpendicular line going through the midpoint of $C_1$ and $C_2$

(iv) **Y-scaling**: scale the distance between $C_3$ and $\overline{C_1C_2}$ to length $b$

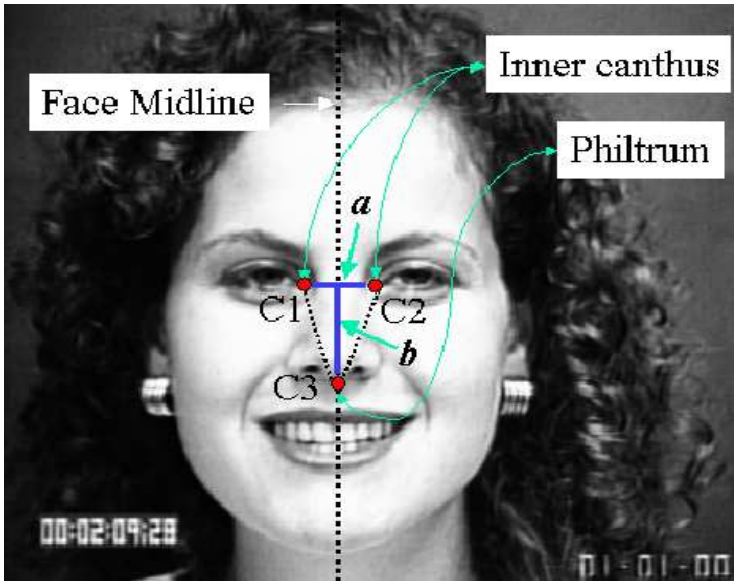


Figure 1: *Face image normalization - $C_1$ and $C_2$ denote the inner canthus of the two eyes, $C_3$ is the philtrum, a is the distance between the points $C_1$ and $C_2$ and b is the distance between the midpoint of $\overline{C_1C_2}$ and $C_3$. The objective of face normalization is to make all of these quantities the same for all subjects in the dataset, which permits a more meaningful comparison. [Courtesy Liu et al, 2003]*

We define the *face midline* as the line going through the midpoint of $\overline{C_1C_2}$ and $C_3$. Each image is cropped into a $128 \times 128$ square image with the face midline centered vertically. The three points $C_1$, $C_2$ and $C_3$ are found in real expression video sequences by hand on the first frame. The rest of the video frames are automatically tracked and then the results are verified and corrected by a human user. The resulting face is an affinely deformed and cropped image of the original, normalized with respect to its own midline. All faces have the inner canthus of each eye and the philtrum in the same pixel locations. In our study, these three points are: $C_1 = [40, 48], C_2 = [88, 48]$ and $C_3 = [64, 84]$, thus $a = 48$ and $b = 36$ (upper left corner has coordinates $[0, 0]$).

Figure 2 shows an example of a video sequence (8 normalized frames) of a subject expressing joy. Figure 3 shows one (normalized) frame for each of the three emotions for 7 subjects, the top

Figure 2: *A subject expressing joy (8 normalized frames).*

panel showing a neutral frame.


Figure 3: *Normalized expressions from* 7 *subjects. Each column represents one subject. Row* 1 *has the frames from neutral expressions. Rows* 2 - 4 *are the peak joy, peak disgust and peak anger expression frames, respectively. [Courtesy Liu et al,* 2003*]*

### 3.2 Facial Asymmetry Measurements

Once a face midline is determined, each point on the normalized face image has a unique corresponding point on the other side of the image. For a normalized face image $I$, a coordinate system defined on the face with the X-axis perpendicular to the face midline and the Y-axis coinciding with the face midline, and its vertically reflected image $I'$ (with respect to Y), we define the following two asymmetry measurements as in Liu et al (2003):

- **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 difference in intensity between the corresponding pixels from the two halves of the face, gives the D-face value at the location $(x, y)$.

- **Edge Orientation Symmetry**: (*S-face*) Edges refer to discontinuities or abrupt changes in the gray-scale intensities of an image. An "edged" image $I_e$ is obtained by running a standard edge detection algorithm on a normalized face image $I$. Let $I'_e$ denote the vertically reflected image of $I_e$. If $\phi$ denotes the angle representing the difference in the orientations of the edges on the two sides of a face, pixel by pixel, we define the S-face measure for the coordinate location $(x, y)$ as

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

Since cosine is an even function, this measure is invariant to the relative orientation of the edges with respect to each other. The two corresponding pixels on the two sides of a face thus have the same S-face value.

These two asymmetry measures capture facial asymmetry from different perspectives; *D-face* is affected by the relative intensity variations from the two halves of the face, while *S-face* is affected by the occurrence of edges. The higher the value of the D-face at a pixel, the more *asymmetrical* that point on the face is, and the higher the value of the S-face at a pixel, the more *symmetrical* that particular point on the face is. Figure 4 shows a normalized face, a D-face and an S-face for three subjects in our dataset.

The two sides of the D-face are exactly opposite of each other (values have the same magnitude but different signs), whereas the S-face is the same on the two sides. So, one half of the face

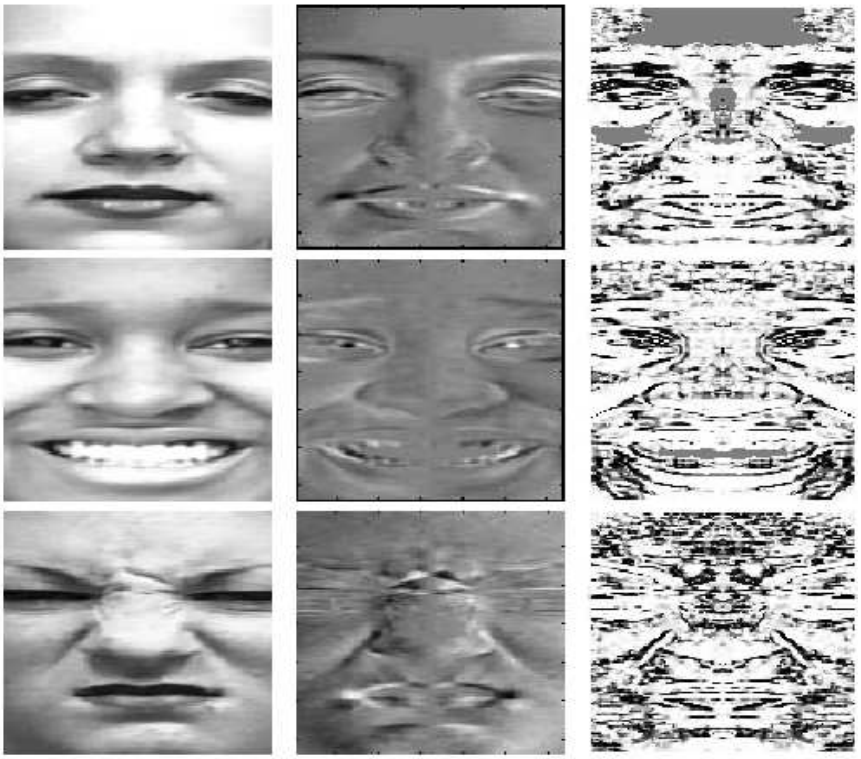

Figure 4: *A normalized face, a D-face and an S-face respectively, in each row, from left to right.[Courtesy Liu et al,* 2003*]*

contains all the information and it is sufficient to consider only this. We average both D-face and S-face values separately over the 128 rows in the half-face to yield the *X-axis values* and also over the 64 columns to yield the corresponding *Y-axis values*, referred to as the "projections". Each element of the Y-axis vector corresponds to a horizontal pixel line in the face, with the first one representing the top of the forehead and the last one representing the bottom of the chin. For the X-axis vector, each element corresponds to a vertical pixel line in the face going from near the ear to the middle of the face. We call these X- and Y-axis elements, the *Asymmetry Faces*. Figure 5 shows the average D-face and average S-face from the 55 subjects, and their most asymmetrical and symmetrical regions. We consider the absolute values of the asymmetry measurements and scale the data for each image frame into the interval $[0, 1]$.

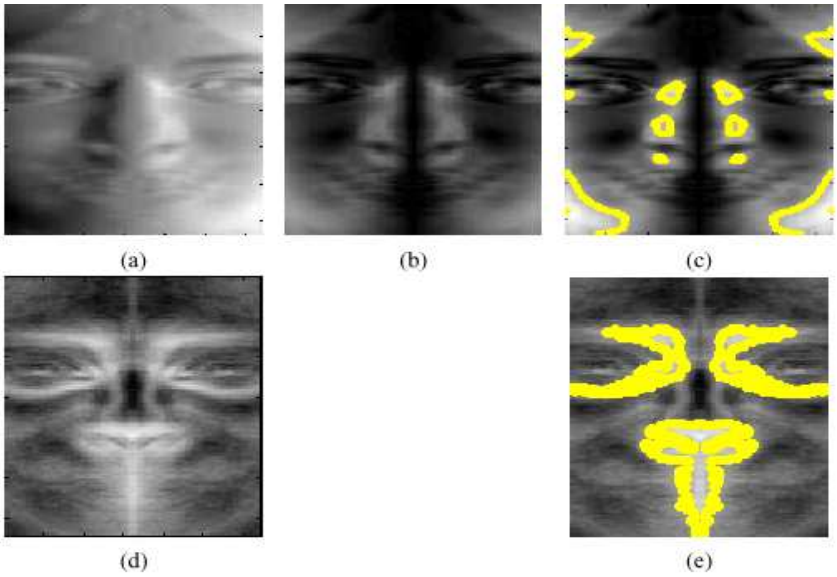


Figure 5: *(a) Average D-face, (b) Absolute values of D-face, (c) the top* 1/3 *most asymmetrical regions by D-face measures are colored yellow, (d) Average S-face, (e) the top* 1/3 *most symmetrical regions by S-face measures are colored yellow. [Courtesy Liu et al,* 2003*]*

Each of the Asymmetry Faces has a total of 8192 ($128 \times 64$) features in all. In order to reduce the cost of computation and to increase classification accuracy by removing redundant features, we perform Principal Component Analysis (PCA). This is done on the $8192 \times 495$ matrix separately for D and S-faces. After examining the eigenvalues, the top 60 principal components are kept for D-face and the top 100 for S-face, as they are able to explain more than 99% of the inherent variations in the two datasets. We also work with these PCs as features, apart from the original asymmetry measurements.

### 3.3 Fisher Faces

A commonly used human identification algorithm in computer vision is *Fisher Faces* (Belhumeur et al. 1997). This method uses a combination of PCA and LDA, producing very clear separation boundaries for the classes in a low dimensional subspace so as to aid in the recognition process, especially under illumination and expression changes. The top 25 Fisher Faces are computed from

the original dataset of $128 \times 64$ features to use as a benchmark for assessing the results based on our asymmetry faces.

Table 1 summarizes the various datasets, the notations used and the number of features contained in each. We deal with the last seven of these datasets in the rest of the paper.

| Notation | Definition | Dimension (features) |
|---|---|---|
| D-face | Intensity Difference Image | $128 \times 128$ |
| S-face | Edge Orientation Symmetry Image | $128 \times 128$ |
| $D_h$ | Left half of D-face | $128 \times 64$ |
| $S_h$ | Left half of S-face | $128 \times 64$ |
| $D_{hx}$ | Column mean of $D_h$ on X | $1 \times 64$ |
| $D_{hy}$ | Row mean of $D_h$ on Y | $128 \times 1$ |
| $S_{hx}$ | Column mean of $S_h$ on X | $1 \times 64$ |
| $S_{hy}$ | Row mean of $S_h$ on Y | $128 \times 1$ |
| D | D-face Principal Components | $60 \times 1$ |
| S | S-face Principal Components | $100 \times 1$ |
| FF | Fisher Faces | $25 \times 1$ |

Table 1: *Descriptions of the different feature sets.* $D_{hx}$, $D_{hy}$ *are D-face feature sets,* $S_{hx}$, $S_{hy}$ *are S-face feature sets. These four are the original Asymmetry Faces. D and S are the PCs derived from the D and S-face feature sets.*

## 4 Experimental Setup

We are primarily interested in two types of classification schemes, giving rise to five different kinds of experimental setup:

(i) Train on all frames from two emotions from the 55 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 the peak frames from all the three emotions of the 55 subjects and test on the neutral ones, and vice versa.

We apply our methods to each of the six Asymmetry Face feature sets, then consider various combinations of these, followed by the Fisher Faces and combinations of the Fisher Faces with the Asymmetry Faces. We begin with our baseline method of LDA with AVR and then apply the two resampling techniques (bagging and RSM) to this baseline method.

# 5 Results

## 5.1 Baseline Method - LDA with AVR

Results from applying LDA to the six Asymmetry Faces are shown in Table 2. The misclassification error for each testing subset is defined as the proportion of cases in the test set that are incorrectly classified out of the total number of cases in that test set. These results show that the D-face features are more effective than the S-face features for expression-invariant classification. Both the D and S-face principal components achieve considerable improvement in the error rates. With the D-face principal components, we have an error rate of 3.03% for testing on both the neutral and the peak frames, which implies that only 5 out of 165 images are misclassified. When comparing the five experiments, we find that the neutral and peak frames are easier to classify than the frames from the three emotions. Of the emotions again, the frames expressing "joy" are the hardest to classify. One possible explanation for this is that the training in this case is done on the frames for anger and disgust, which have similar facial expressions (e.g. downturned mouth) and are different

from joy (upturned mouth).

| 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% |

Table 2: **Misclassification error rates** *for the Asymmetry Face feature sets - using the baseline method of LDA with AVR.*

Table 3 shows results from combining some of the Asymmetry Face feature sets. We see that adding on more features gradually improves the classification. This is not unexpected, as we hope to gain strength from having more information. Figure 6 shows the decrease in the error rates as the number of combined feature sets increases. Perfect classification results (0% error rate) are obtained for testing on the neutral and the peak frames with all six Asymmetry Faces together. For the emotions, only 2 frames for anger, 5 for disgust and 11 for joy are misclassified (out of a total of 165 in each case).

The goal of the experiments with the Fisher Faces and their combination with the Asymmetry Faces is to assess the performance of the Asymmetry Faces vis à vis the Fisher Faces, and to investigate whether Fisher Face and Asymmetry Face components provide complementary information for better discrimination of human faces under expression variations. For the combinations, the Fisher Faces are concatenated with the Asymmetry Faces and then the classifiers are applied to the enlarged feature space. Table 4 gives some of these results. They indicate that the misclassification rates get lower and lower as more and more Asymmetry Faces are combined with the

| Feature Set/Test Set | joy | anger | disgust | neutral | peak |
|---|---|---|---|---|---|
| $D_{hy}$+$S_{hy}$ | 19.39% | 12.12% | 22.42% | 6.67% | 9.70% |
| $D_{hx}$+$S_{hx}$ | 48.48% | 43.64% | 48.48% | 33.94% | 36.97% |
| 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% |

Table 3: **Misclassification error rates** *from combinations of Asymmetry Faces.*

Fisher Faces, which is again demonstrated by Figure 7. For example, the error rate for testing on joy using only the Fisher Faces is 12.12%, which decreases to 5.45% by adding on the D-face and the S-face principal components, an improvement of 55%. Similarly, the performance of the Asymmetry Faces improves by adding on the Fisher Face features. For example, error rates for testing on joy for D-face PCs is 18.18% (see Table 2) which is reduced to 8.48% by adding on the Fisher Faces (a reduction of 53.3%). Finally, when all six Asymmetry Face feature sets are added to the Fisher Faces, we obtain perfect classification (0% error rates) for four out of the five testing experiments. This confirms our belief that the Asymmetry Faces and the Fisher Faces are capable of complementing each other and thus produce better results than either of these feature components alone.

The results based on the Asymmetry Faces (shown in Tables 2 and 3) are significantly better than pure random guessing. The probability of misclassification by random guessing in the case of 55 classes is 98%. We carried out a comparison of the misclassification rates obtained from the

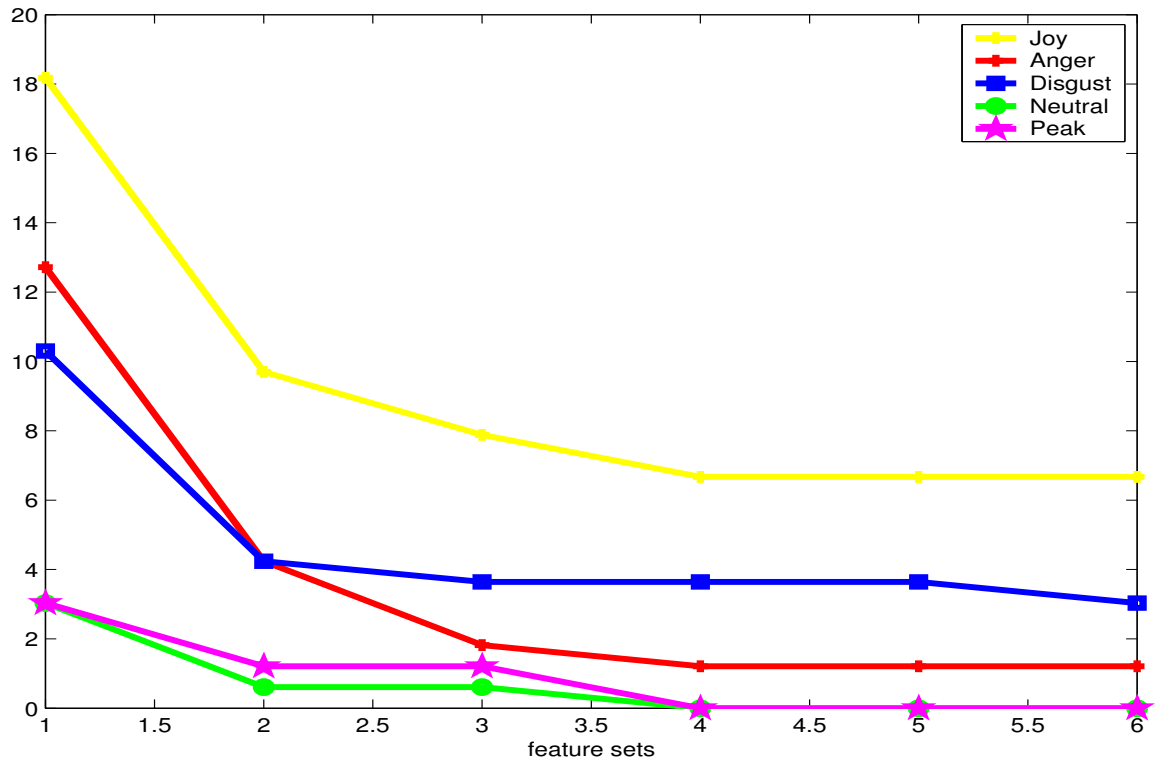


Figure 6: **Misclassification error rates** *for the different orders of combination of the Asymmetry Face feature sets. The labels on the x-axis show the number of feature sets in the respective combinations that produced the lowest error rates. The best combination for joy based on* 6 *feature sets has a misclassification rate of* 6.67%*, which differs from the best for neutral which is* 0%*.*

Asymmetry Faces with the random guessing rate using standard statistical tests for difference in proportions. When comparing the performance of the Fisher Faces with that of the six Asymmetry Faces, we found no statistically significant differences in the error rates for any of the five testing subsets. Statistical tests to gauge the improvement in the Fisher Face results by combination with one or more of the Asymmetry Face feature sets show that for testing on anger and peak frames, both Fisher Faces and Asymmetry Faces already achieve very good results on their own (error rates: 1.82% and 0.61% for Fisher Faces, see Table 4; 1.21% and 0% for Asymmetry Faces, see Table 3), thus significant improvements are not observed. Significant improvements are however observed for testing on joy, disgust and neutral frames.

| Feature Set/Test Set | joy | anger | disgust | neutral | peak |
|---|---|---|---|---|---|
| FF | 12.12% | 1.82% | 4.24% | 3.64% | 0.61% |
| FF+D | 8.48% | 1.82% | 1.82% | 0.00% | 0.00% |
| FF+ S | 6.67% | 1.21% | 2.42% | 1.21% | 0.00% |
| FF+D+S | 5.45% | 0.61% | 1.21% | 0.00% | 0.00% |
| FF+D+S+$D_{hy}$ | 5.45% | 0.61% | 0.00% | 0.00% | 0.00% |
| FF+D+S+$S_{hy}$ | 4.24% | 0.61% | 0.61% | 0.00% | 0.00% |
| FF+D+S+$D_{hy}$+$S_{hy}$ | 4.24% | 0.00% | 0.00% | 0.00% | 0.00% |
| FF+D+S+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 2.42% | 0.61% | 0.61% | 0.00% | 0.00% |
| FF+D+S+$D_{hy}$+$S_{hy}$+$S_{hx}$ | 4.24% | 0.00% | 0.00% | 0.00% | 0.00% |
| FF+D+S+$D_{hy}$+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 2.42% | 0.00% | 0.00% | 0.00% | 0.00% |

Table 4: **Misclassification error rates** *for Fisher Faces, alone and in combination with Asymmetry Face feature sets.*

## 5.2 Results from the Resampling Methods

The primary objective of employing resampling methods is to improve upon the results from the baseline classifier. Recall here that we obtained very good classification results when combining many feature sets together, for example, when combining all the Asymmetry Face feature sets and also when combining all the Asymmetry Face feature sets with the Fisher Faces. However, we expect that bagging and RSM will achieve comparable, or better, results with even fewer feature sets, which will be a definite improvement over the existing results reported in Section 5.1.

### 5.2.1 Bagging

After experimenting with different numbers of resamples (from 10 - 100), we use $60 - 70$ resamples for bagging, as this gave optimal results in most cases. We repeat the entire resampling procedure

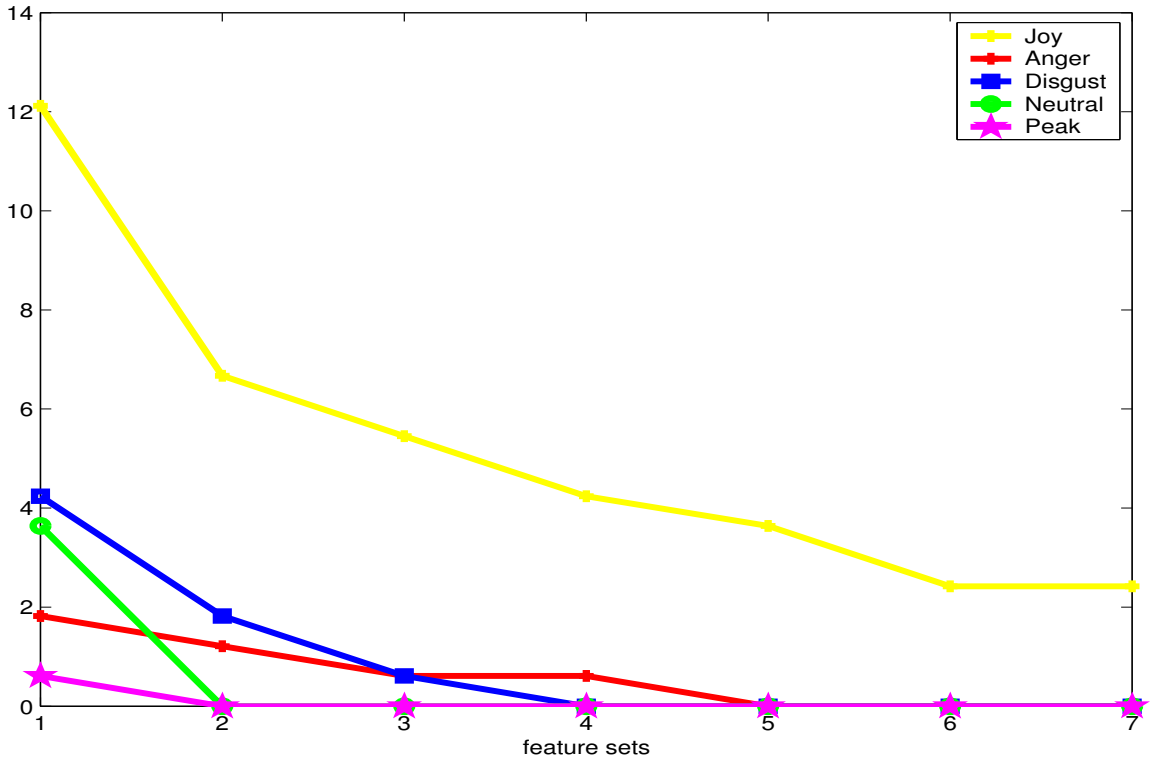


Figure 7: **Misclassification error rates** *for the different orders of combination of the Fisher Faces with the Asymmetry Face feature sets. The labels on the x-axis show the number of feature sets in the respective combinations that produced the lowest error rates. The best combination for joy based on the* 7 *feature sets has a misclassification rate of* 2.42%*, which differs from the best for neutral which is* 0%*.*

20 times and obtain the final misclassification errors by averaging over these 20 iterations. As can be seen in Table 5, an appreciable improvement in the classification performance of all the feature sets results from bagging.

### 5.2.2 RSM

Table 6 shows the improvement rates achieved by RSM over the Asymmetry Face feature sets. In each case, $p^*$ is 50% of $p$, the total number of features available.

RSM is known to perform well when there is some redundancy present in the feature space,

| Feature Set/Test set | joy | anger | disgust | neutral | peak |
|---|---|---|---|---|---|
| $D_{hy}$ | 43.96% | 51.16% | 26.49% | 77.15% | 56.73% |
| $S_{hy}$ | 44.71% | 34.59% | 42.75% | 57.13% | 51.33% |
| $D_{hx}$ | 10.52% | 15.05% | 14.22% | 24.78% | 33.54% |
| $S_{hx}$ | 15.73% | 25.10% | 20.71% | 32.24% | 26.52% |
| D | 38.33% | 59.50% | 45.87% | 74.00% | 47.99% |
| S | 38.89% | 51.62% | 45.33% | 52.51% | 58.06% |
| FF | 45.73% | 65.03% | 69.33% | 30.67% | 85.10% |
| $D_{hy}$+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 52.42% | 72.50% | 53.16% | 77.76% | 92.50% |
| D+S+$D_{hy}$+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 24.13% | 7.44% | 75.00% | - | - |
| FF+D+S+$D_{hy}$+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 89.98% | - | - | - | - |

Table 5: **Improvement rates** *achieved by bagging over the baseline LDA+AVR results for the individual Asymmetry Faces and the Fisher Faces (shown in Tables 2, 3 and 4), computed as (baseline error-bagging error)/baseline error* $\times 100$. *A "-" denotes that bagging was not applied to those test sets since they had perfect classification (*0% *error rate) with the baseline classifier.*

otherwise, it is not guaranteed to give good results (Skurichina and Duin, 2001). $D_{hy}$, $S_{hy}$, $D_{hx}$, $S_{hx}$ and their combination have considerable redundancy and RSM is found to improve over the baseline LDA results. The principal components 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. Indeed, there is either a deterioration or only a marginal improvement over the original results for D and S (not shown here).

| | joy | anger | disgust | neutral | peak |
|---|---|---|---|---|---|
| $D_{hy}$ | 31.46 | 35.83 | 13.42 | 62.39 | 34.15 |
| $S_{hy}$ | 30.28 | 22.79 | 26.76 | 39.04 | 25.52 |
| $D_{hx}$ | 6.09 | 3.11 | 8.69 | 16.06 | 6.83 |
| $S_{hx}$ | 3.46 | 9.24 | 10.35 | 20.43 | 9.27 |
| $D_{hy}$+$S_{hy}$+$D_{hx}$+$S_{hx}$ | 39.94 | 52.78 | 42.66 | 77.76 | 74.57 |

Table 6: **Improvement rates (%)** *over the baseline LDA+AVR results from applying RSM to the Asymmetry Faces (shown in Tables 2 and 3), computed as (baseline error-RSM error)/baseline error* $\times 100$.

### 5.3 Comparison of Resampling Results

Bagging results are significantly better than those from the baseline classifier for all the four original asymmetry face feature sets. For the PCs and the Fisher Faces, they are not significantly better for a few of the test sets, which might be attributed to the fact that the original errors for these are quite low to start with and hence there is not much scope for improvement. When comparing with the baseline results, it is possible to attain improvement rates anywhere between $10-90\%$. For the combinations of the feature sets, there is significant improvement in all cases. It thus proved to be worthwhile to apply bagging to most of our datasets.

RSM does not achieve improvement over the original results as uniformly as bagging. Only for the S-face Y-projections and the combination of the four original Asymmetry Faces, do we observe statistically significant improvements for all the five testing subsets. Significant improvements for the other datasets are limited. However, statistical tests reveal no significant differences between the bagging and the RSM results in most of the cases. Table 7 gives a summary of the cases for which bagging produced significantly better results than RSM at the 5% level of significance.

Apart from the amount of improvement, another perspective from which bagging and RSM

| | |
|---|---|
| $D_{hy}$ | peak |
| $S_{hy}$ | digust, peak |
| $D_{hx}$ | peak |
| $S_{hx}$ | joy, anger, peak |

Table 7: *Test sets for which bagging performed significantly better than RSM.*

could be compared is their scope of application. RSM is heavily dependent on the nature of the features themselves and the way they interact with each other, and often deteriorates performance when some feature extraction has already been performed. Many computer vision problems are high-dimensional with irrelevant and redundant features. Dimension-reduction is usually employed in these cases. This highlights a limitation of the RSM procedure. Bagging does not suffer from this constraint, because it is not influenced by the nature of the features (since they are all retained); all that matters is whether the particular classifier being used is unstable or not for that specific dataset, which is not too difficult to determine. From all these perspectives, it is evident that bagging is a more flexible procedure with much wider applicability.

## 6 Selected Feature Subsets

Feature selection is a major ingredient in any human face recognition algorithm. It enables us to identify the particular features that help in the recognition process. Recall that we did feature selection depending on the AVR values of the features. The higher the AVR value of a feature, the more discriminating power it possesses and the more crucial it is for identification purposes. Figure 8 shows AVR values of the four original asymmetry face feature sets for subject identification. The values indicate that features in and around the nose bridge and some in the forehead region play the most prominent role in human identification. These features are marked in Figure 9.

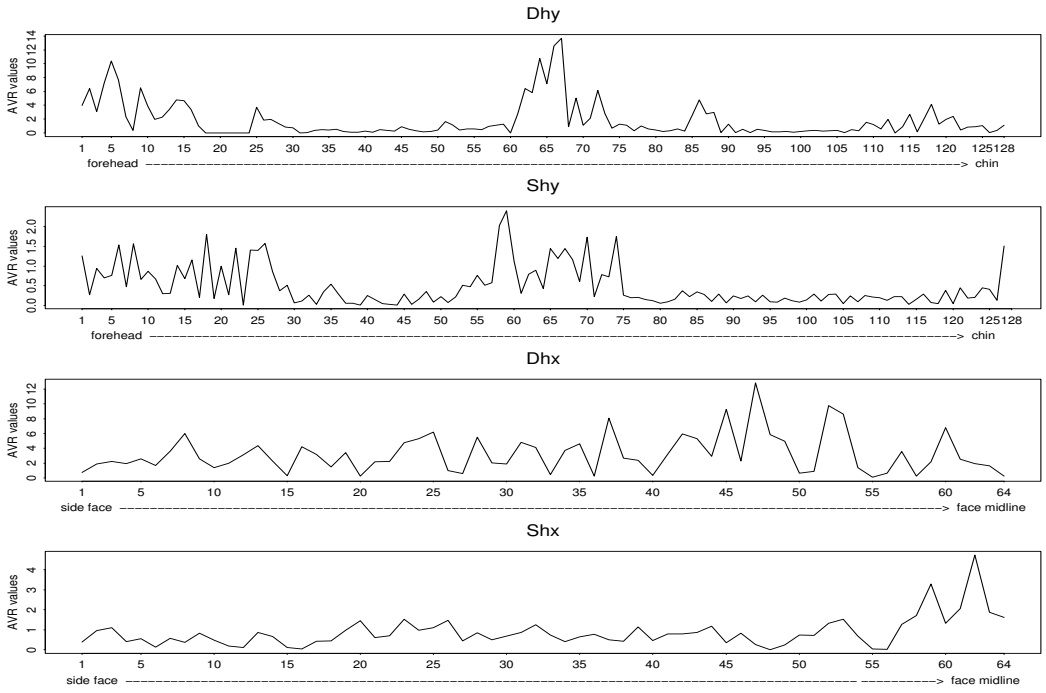


Figure 8: *AVR values for the features from the four Asymmetry Faces for subject classifications. Features around the nose and some around the forehead have high AVR values, hence prove to be discriminatory across the* 55 *subjects in our study.*

## 6.1 Distinguishing Males and Females

Another interesting observation from studying the feature sets is the presence of ample evidence for the existence of statistically significant differences in the asymmetry feature values for the two sexes. More specifically, males possess a significantly higher degree of facial asymmetry than females (higher D-face values but lower S-face values). The particular facial features that differ markedly in the amount of asymmetry between males and females are mostly around the mouth, chin and above the eyes. The AVR plots for discriminating males from females are in Figure 10; the features with the highest AVR values are marked in Figure 11. The overall impression from these analyses is that there is an appreciable difference in the quantified facial asymmetry information for males and females over a large region of the face. This agrees with results reported in Liu and Palmer (2003) on differences in the asymmetry measures between the two sexes for 3-D human face images.

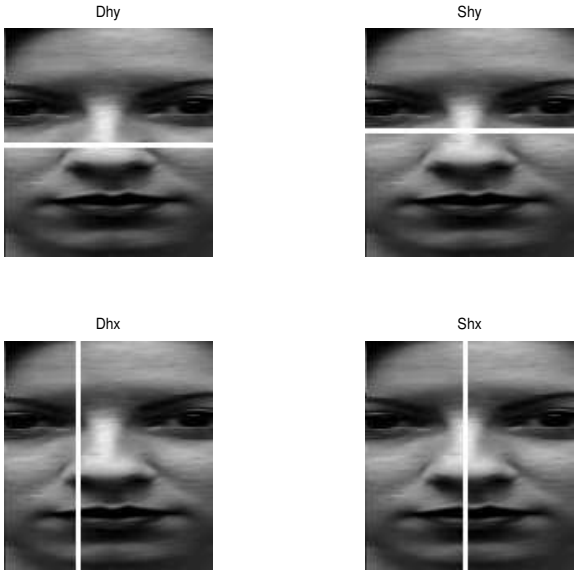


Figure 9: *The facial features with the highest AVR values for human identification. The white lines correspond to those facial features that have the highest AVR values as shown in Figure 8.*

## 6.2 Human vs. Expression Identification

The final observation from the exploratory analysis with the feature subsets is that different subsets of features help in human identification and expression identification, that is, the asymmetry of different parts of the face help in telling individuals apart (eg, Tom from Harry) and expressions apart (eg, joy from anger). Figure 12 shows the AVR values of the features for expression classification. The features with the highest AVR values are marked in Figure 13. A quick comparison with Figure 9 reveals that unlike the case of human identification where the bridge of the nose turned out to be the most discriminating, the region just above the mouth is discriminating across different expressions. We see that the asymmetry of different parts of the face plays a role in expression identification and expression-invariant human identification.

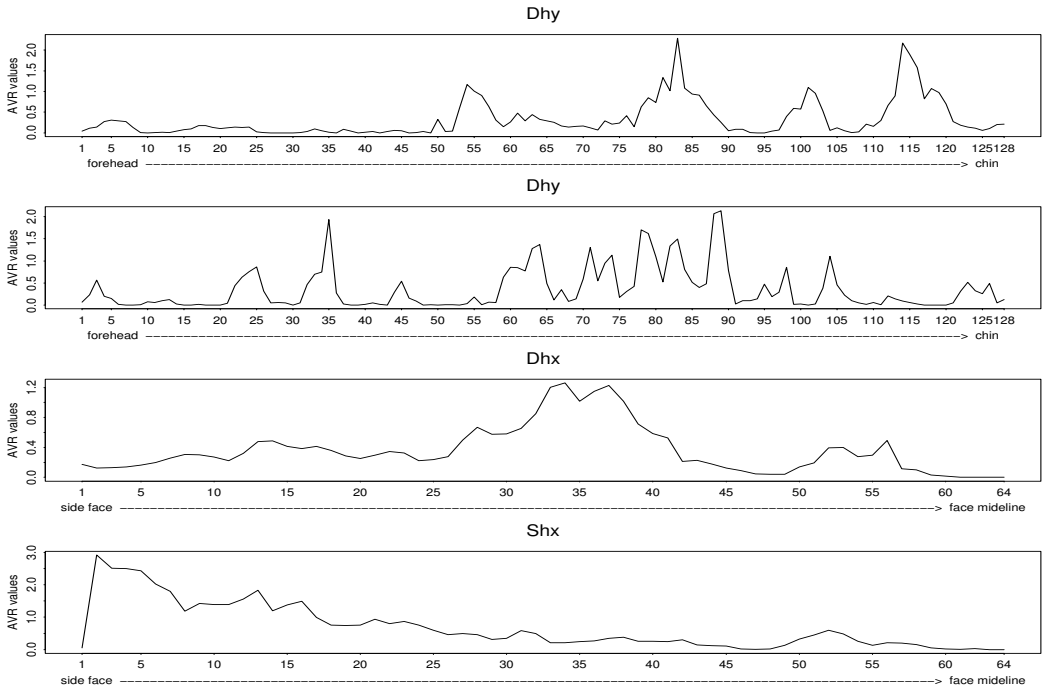


Figure 10: *AVR values for the different features for sex identification. Regions around the mouth, chin, above the eyes appear to be the most discriminatory for males and females.*

# 7 Discussion

As also reported in Liu et al (2003), the baseline method of LDA gives satisfactory results for this 55-class problem. The results are encouraging indeed, given the large dimension of our problem, and indicate that asymmetry measures are quite effective as features for recognizing people under expression variations. Their performance is comparable with that of Fisher Faces, a popular human identification algorithm. Combining Asymmetry Faces and Fisher Faces leads to significant improvement in classification rates. Asymmetry features are simple to compute and this fact gives them an advantage, namely, a wide scope of applicability in designing automatic face recognition systems for handling large real databases that arise from surveillance cameras at airports, ATMs and other public places.

The baseline method is further improved upon by the resampling methods. In particular,

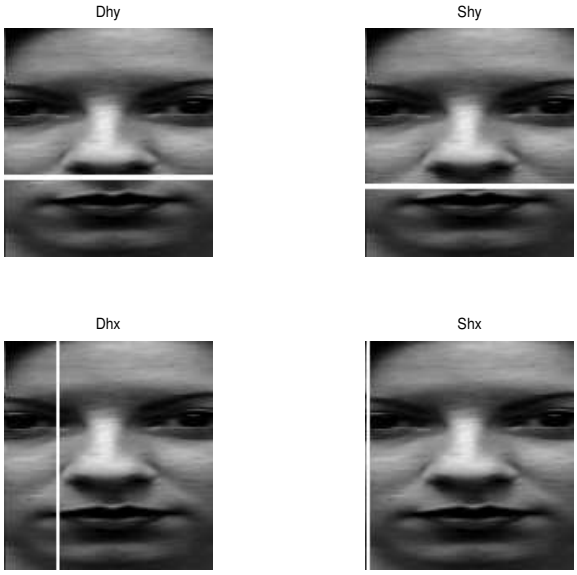


Figure 11: *The facial features with the highest AVR values for sex identification. The white lines correspond to those facial features that have the highest AVR values as shown in Figure 10.*

bagging is able to produce near perfect classification in many cases. We have thus established that it is possible to achieve excellent classification results with very little additional resources. In fact, we have shown that the resampling methods enable us to achieve very good classification results with fewer feature set combinations. For example, the results that we obtained with four Asymmetry Faces together with bagging and RSM are not significantly different from those with six Asymmtery Faces using the baseline method of LDA with AVR. Furthermore, the ease of their implementation also makes such methods useful from a practical point of view, since most real-life applications require quick, accurate algorithms. Therefore, the use of resampling methods is expected to be widely applicable.

As mentioned earlier, both bagging and RSM are primarily based on bootstrapping although in semantically different ways - one samples from the training sets while the other samples from the features. Whereas bagging is a standard application of case-resampling bootstrap, RSM is a rather unconventional application. We do not usually think of bootstrapping variables in a model

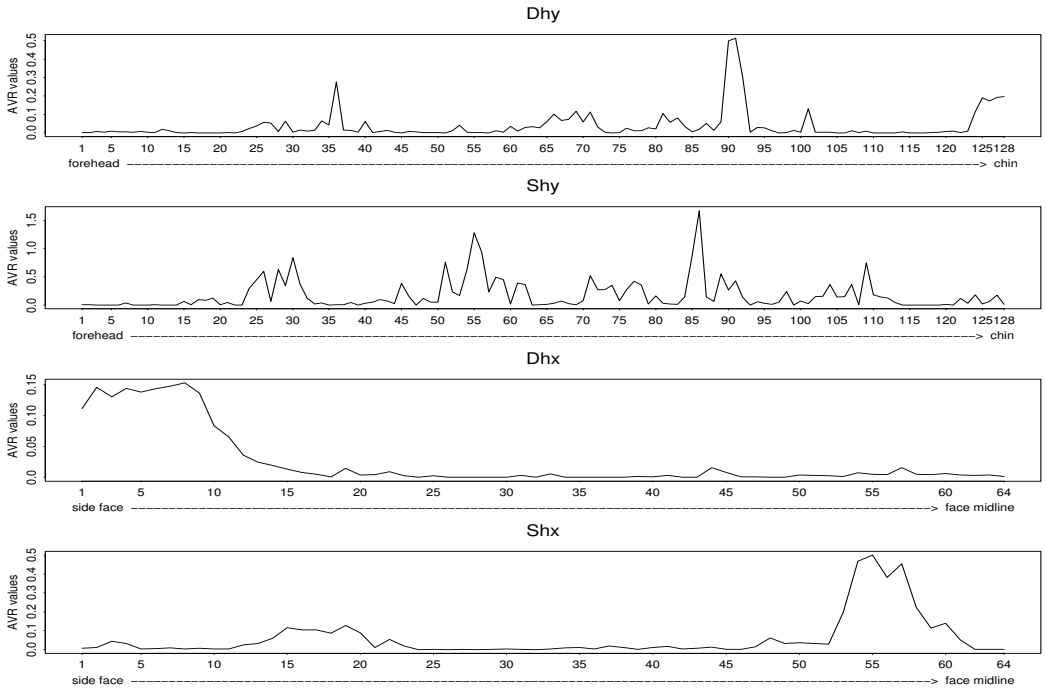


Figure 12: *AVR values for the different features for expression identification. We have three different types of expressions - joy, anger and disgust. Regions around the mouth and side of the face appear to be the most discriminatory across emotions.*

selection process, although there is no impediment to doing so. Even though the approaches differ, the goal is the same - to improve upon the results of the earlier baseline method.

It is also worth recalling here another well-known resampling technique known as boosting (Freund and Schapire, 1997) which has been found to work well in a variety of pattern recognition problems. But according to Skurichina and Duin (2000), boosting does not depend on the instability of the base 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. Our initial experiments with this classifier did not give very good results. We hope to further investigate whether modifications of the original boosting algorithm will obviate these problems.

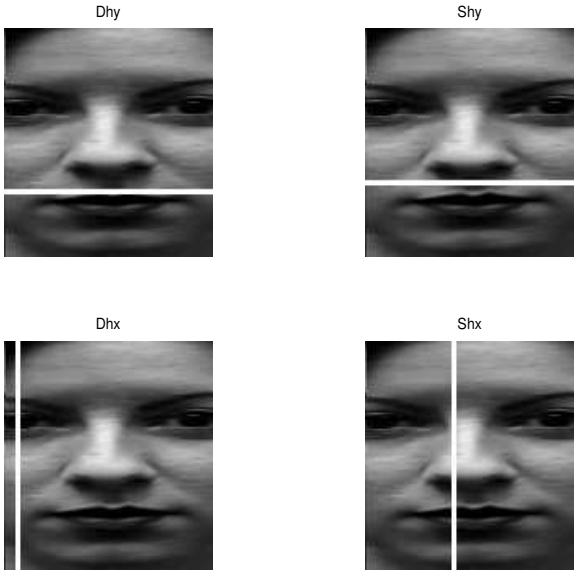


Figure 13: *The facial features with the highest AVR values for expression identification. The white lines correspond to those facial features that have the highest AVR values as shown in Figure 12.*

At this point, a reader might wonder whether our results will be valid also for expressions that are produced spontaneously, since the people in the current 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 image with a spontaneous expression. Indeed, according to Hager and Ekman (1985), posed facial expressions are more asymmetric than genuine and spontaneous ones, and thus our current methods might be providing a biased estimate of the extent to which facial asymmetry can truly aid in face recognition. It would be interesting to investigate how the classification results change when using genuine expressions.

One thing we wish to point out at the end is that our dataset comprises university students only and hence is not representative of the general population, especially in terms of age. For example, the group of people that passes through an airport or accesses the ATM is more diversified with respect to age, ethnicity and certain other demographic factors than the one that attends a

particular college or university. Although we expect our techniques to yield fairly good results on larger and more diverse databases, we intend to explore this issue for assessing the generalizability of our technique. This would further enhance the utility of our methods in real-life applications.

## Acknowledgement

The authors wish to thank Jeff Cohn and Karen Schmidt for providing the database.

## References

[1] Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, second edition.

[2] 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.

[3] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

[4] 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.

[5] Hager, J. and Ekman, P. (1985). The asymmetry of facial actions is inconsistent with models of hemispheric specialization. *Psychophysiology*, 22:307–318.

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

[7] 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. Publicly available at $< http://www.ri.cmu.edu/projects/project\_420.html >$.

[8] 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*.

[9] 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.

[10] 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)*.

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

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

[13] Skurichina, M. and Duin, R P W. (2000). Boosting in linear discriminant analysis. In *Lecture Notes in Computer Science*, volume 1857. Springer-Verlag, Berlin.

[14] Skurichina, M. and Duin, R. P. W. (2001). Bagging and the random subspace method for redundant feature spaces. In *Lecture Notes in Computer Science*, volume 2096. Springer-Verlag, Berlin.

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

[16] 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.