Benchmark Studies on Face Recognition

Srinivas Gutta, Jeffrey Huang, Dig Singh, Imran Shah,
Barnabás Takács, and Harry Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030
USA

Abstract

We describe herein FERET - a major R&D program aimed at advancing the state-of-the-art in automated face recognition. We address three complementary problems: (A) development of a representative data base set of facial images to train, test, and evaluate alternative face recognition schemes; (B) bench marking of both simple but well known algorithms, and of novel automated and integrated face recognition schemes; and (C) training and testing of human subjects to evaluate human performance using the same data base developed to test the automated face recognition component. The major results of our R&D program indicate that (i) future advances in automated face recognition are predicated on the development of hybrid recognition systems, (ii) that holistic (connectionist) methods outperform discrete (and direct) feature and correlation methods, (iii) that if the test beds are random and contextual cues are lacking human performance is quite poor, not consistent, and degrades rapidly when compared with machine performance, and (iv) that extensive and proper testing is crucial for bench marking.

1 Introduction

FERET, a major R&D program funded by ARPA and ARL, and described herein aims at advancing the state-of-the art in automated face recognition. The specific problem under consideration, known as the Face in the Crowd, is that of automatically identifying a presentation of a face as being one of many resident in a data base of previously stored faces. As the Interface Design Document (IDD) specified under this research program involves an active data gathering sensor there is the possibility to request another look at the face and then to fuse the data for enhanced face identification. We address three complementary problems: (A) development of a representative data base set of facial images to train, test, and evaluate alternative face recognition schemes; (B) bench marking of both simple but well known algorithms, and of novel automated and integrated face recognition schemes; and (C) training and testing of human subjects to evaluate human performance using the same data base developed to test the automated face recognition component.

The face is a unique feature of human beings. Even the faces of "identical twins" differ in some respects. The uniqueness of faces is also the reason for their widespread use in applications where identification of people is important. Humans detect and identify faces in a scene with little or no effort. However, building automated systems that accomplish this task is very difficult. There are several related sub problems: detection of a pattern as a face, identification of the face, analysis of facial expressions, and classification based on physical features of the face (Samal and Iyengar, 1992). A system that performs these operations will find many applications, e.g. criminal identification and retrieval of missing children. The specific problem under investigation herein is that of classifying an image cropped and detailed using reckoning and foveation.

Face recognition is difficult mostly because of the inherent variability of the image formation process in terms of image quality, geometry, and/or occlusion, change, and disguise. All face recognition systems available today can perform only on restricted data bases of images in terms of size, age, gender, and/or race, and they further assume well controlled environments. The Face in the Crowd problem would involve, however, several hundred and possibly thousands subjects, of varying age, gender, and race, to be matched against a relatively limited data set (10-20) of prestored facial images. There are different possible degrees of variability ranging from those assuming that the position/cropping of the face and its environment (distance and illumination) are totally controlled, to those involving little or no control over the background and viewpoint, and eventually to those allowing for major changes in facial appearance due to factors such as aging and disguise (hat and/or glasses).

The inherent variability characteristic of the image formation process has to be fully reflected during the development of the data base of facial images. One needs to provide adequate variability and sufficient test images in order to perform meaningful bench marking and statistical evaluation of alternative automated face recognition systems and human performance. Image variability has to be expressed along several dimensions such as face characteristics (age, gender, and race), geometry (distance and viewpoint), image quality (resolution, illumination, and signal to noise ratio), and image contents (occlusion and disguise). Statistical evaluation

involves cross validation (CV) using disjoint sets of training, tuning, and test data obtained under similar conditions of image formation. As already stated earlier, machine and human performance evaluation are complementary and should benefit each other for further refinement.

The basic schemes for face recognition implemented and evaluated using the FERET data base include methods representative of the two major approaches for automated face identification. The abstractive (feature point extraction) approach seeks to define a set of key parameters for the measurement of faces, and to subsequently employ standard statistical pattern recognition techniques for matching amongst faces using these measurements. To automate such measurements involves finding ways to locate facial features within images - a problem which has proved remarkably difficult. In contrast to the abstractive approach, one can use holistic (multilayered connectionist) approaches characteristic of methods such as backpropagation and projection pursuit, principal component analysis (PCA) and singular value decomposition (SVD) in terms of eigenfaces, synthetic discriminant functions (SDFs), and radial basis functions (RBFs). Note that the holistic approach still involves deriving appropriate measurements, but the task is easier now because the hidden nodes of the multilayer architectures can define and measure abstract features.

2. Facial Database Collection

Specific software and hardware requirements have to be met in order to accomplish the goals of our R&D program in terms of data collection, communication and sharing of resources, data processing and bench marking, and graphical user interfaces (GUIs) and human benchmark studies. Computational requirements are extensive, even more so when one contemplates processing of hundreds of images at the same time.

2.1 Lab Configuration

Hardware consists now of SPARC 20 (with 19" color monitor) and LX stations (operating under Solaris 2.3 OS), Apple Quadra 840 AV and several Apple Quadra 660 AV with 16" color monitors (operating under Apple System 7.1). Image acquisition is further facilitated by the availability of 8mm SONY videoHi8 Handycam camcorder and Video Cassette Recorder SONY SLV-R1000. Archiving, communication, and information exchange are facilitated by 7GB SCSI Disk, 1/4" tape drive, and external 644 MB CD-ROM for the SUNs, 1GB and 500 MB SCSI hard drives, 4mm 2GB DAT DDS tape, and CD 300i External CD-ROM for the Apple-Quadras, and ftp and internet communications. Utilities and development software include amongst other things Adobe PhotoShop Deluxe, MATLAB 4.1 and corresponding image, signal and neural network processing toolboxes.

2.2 Data Collection

755 facial image sets - including 113 duplicate sets taken at different times and possibly wearing glasses - consisting of several poses and totaling 5,045 images have been collected so far. Most of the sets consist of the following poses: two frontal shots ('fa' and 'fb'), 1/4 half (right and left) profiles ('qr' and 'ql'), 3/4 half (right and left) profiles ('hr' and 'hl'), and right and left (90 deg.) profiles ('pr' and 'pl'). In addition we recently collected several hundred sets that have several additional poses at the midpoints between: 'hr' and 'qr' ('ra'), 'qr' and frontal view ('rb'), frontal view and 'ql'('rc'), 'ql' and 'hl' ('rd'), and between 'hl' and 'pl' ('rd'). The additional poses were taken to assess the capability of modeling the human face using several positions and interpolating /extrapolating amongst them for identification purposes.

Most of the facial image sets were acquired using Kodak Gold Ultra 400 film and developed on CD - ROM using five resolutions: High Scan 16 (2048 x 3072), High Scan 4 (1024 x 1536), High Scan 0 (512 x 768), High Scan -4 (256 x 384), and High Scan -16 (128 x 192). KODAK CD-ROM format is compatible with both SUN and Apple - QUADRA AV series. KODAK CD-ROMs come with (hard copy) thumbnail prints of each set of images to facilitate manual browsing.

3. Methodology

The methodology followed in our R&D program is predicated on (i) the tasks we set to solve and (ii) the concept of integrated and hybrid recognition systems. In terms of data availability one can assume that, initially at least, segmented ('cropped') and normalized face image sets would become available as part of the database acquisition effort followed by relatively straightforward procedures for face detection and boxing. Note that the facial expressions for each facial image could possibly vary. During recognition, information could be possibly be made available regarding the geometry ('pose') of the face image, but the pose being tested does not necessarily belong to the training set. There are two possible recognition tasks to be considered under this program.

MATCH - Identify a large number of images, possibly in excess of one thousand, against a database (DB) of the same size. The major goals here are to assess robustness and how identification scales up with increasing the size of the DB.

SURVEILLANCE - Check if some facial image belongs to a relatively small but predefined set of images (DB#1). The goal here is to assess how the identification ('recognition') system performs when 'flooded' with possibly hundreds of false positives. The size of the database (DB#2), including mostly

false positives but also some images belonging to DB#1, is one or two order of magnitudes larger than the size of DB#1. One possible scenario would have about 10-20 images in DB#1 and about 100-200 in DB#2.

Testing procedures have to be established to assess performance (accuracy and reaction time) and learning capabilities like incremental training. The image database can be partitioned into three sets: development - A, training and tuning - B, and testing - C. As label ('tag') information is available on each facial image additional postmortem analysis of performance as a function of ethnicity, gender, age, and pose becomes possible.

Machine benchmark studies were done using variants of both the connectionist and discrete approach. The *Radial Basis Functions* (RBFs) paradigm has been the main holistic approach used due to (i) its conceptual similarity to other neural paradigms, (ii) integrated (unsupervised and supervised) nature, (iii) relatively simplicity in implementation and evaluation. Learning involves here a hybrid systems consisting of (a) unsupervised training of the hidden layer using some standard clustering (k - means) algorithm to define prototype face images (kernel basis functions / PCA-like); and (ii) supervised training of the output classification layer using correlation matrices based (2nd order statistics) LMS. The *Feature (Discrete) and Correlation Approach* (FCA) paradigm involves detection of characteristic facial features like the eyes, nose, mouth, etc, and measuring their relationships. Inflection points, possibly using scale-space methods or 'manual' measurements, are used to derive appropriate feature vectors which are then fed into simple classifiers such as k - nearest neighbors or correlated possibly using affine transformations.

Human benchmark were carried out in parallel to the machine benchmark studies to provide comparative performance. Recognition time and accuracy should be measured as a function of (i) size of the database for the MATCH task or the relative sizes of the databases involved in the SURVEILLANCE task, (ii) display and latency times, and (iii) image quality.

4. Human Benchmark Studies

Both the human benchmark studies and the Feature (discrete) and Correlation (FCA) approach require graphical image presentation and interface via properly designed Graphical (friendly) User Interfaces (GUIs).

4.1 Graphical User Interfaces (GUI)

The specific GUI designed for our project was done using X11/Interviews for X-windows. The tool kit developed involves user interface ('command buttons') and specific applications, and it is supported through the Interviews library by Xlib and Xserver. The learning interface amounts to a picture browser so the users can be trained on images of human faces. During testing, face images are displayed for some predefined time, and human subjects are prompted to make their choice (YES, NO, maybe with some CONFIDENCE ranging from 0 to 10) with respect to the image presented for their inspection. The same GUI also facilitates probing face images and collecting data - determining locations of features related to facial landmarks such as the eyebrows, eye, nose, mouth, and chin - as required by the Feature (Discrete) and Correlation approach.

4.2 Delay and Latency Human Surveillance Studies

The human benchmark studies tested mostly on performance accuracy and on how performance degrades as the time between training and testing increases. In one of the experiments carried out, two control group were used, each consisting of 30 subjects. Each group was trained on the SURVEILLANCE task by being shown 10 images while during testing the group would be tested on 40 images, with only some of them coming from the original training set. Group 1 was tested two hours after training while Group 2 was tested one week after training. The images in the test set were chosen so there are specific differences between them and those in the training set with respect to presence/absence of spectacles and/or smile. The results clearly indicate that humans perform quite poorly when tested on different exposures from those originally trained on and that, as one would expect, performance degrades as the time interval between training and testing increases. One possible remedial training activity will possibly require repetitive training and also training with several exposures of the same subject so temporal differences with respect to noise and geometrical changes are properly captured.

5. Machine Benchmark Studies

Face recognition usually starts through the detection of a pattern as a face, proceeds then by normalizing the face image to account for geometry and illumination changes using information about the box surrounding the face and/or the eye location, and finally identifies the face using appropriate image representation and classification algorithms.

5.1 Face Detection

A novel dynamic vision scheme based on a biologically motivated feature extraction model involving retinal sampling along M-type lattices ('where' channel) and small oscillatory eye movements has been developed (Takacs and Wechsler, 1994). The bounding box of the face is then found based on the conspicuity surface emerging

while scanning the input and deriving the features. Simulation results on over 400 images showed an error rate of less than 5%.

5.2 Eye Detection

The process of eye detection is complementary to that of face detection. It also involves retinal sampling performed this time along P-type lattices ('what' channel) and microsaccades, while classification is done using the self-organizing feature map (SOFM). The optimal set of eye templates is found by an enhanced SOFM approach using cross validation training (Takacs and Wechsler, 1994). The approach was tested on 50 images including faces from both the training set and the test sets (the noise set was represented by non-eye pixels). The algorithm reliably marked the eye regions on all of the tested faces, including images where the subjects wear glasses. As some false positives detection also occurred additional model based post processing is required.

5.2. Holistic Approach

The best results on the SURVEILLANCE task were obtained using Ensemble networks consisting of RBFs and trained on both original and distorted facial images. The ERBF where trained over the original image and slightly degraded (by Gaussian noise, blur, and/or geometrical transformations - small translation and rotation) corresponding images. The ERBF itself consists of three RBF networks, each slightly different from the other in terms of the number of clusters and overlap factors, and the outputs of each network are normalized so they define a pdf (probability density function). The final decision as to whether to accept an image or to reject it is based on the following rule:

If the norm of the average of all (nine) outputs is greater than 0.60 then accept ('recognize') else reject;

Using the above architecture we tested ERBF using cross validation on a set consisting of 500 frontal 'fa' and 'fb' images. Using cross validation (CV) one alternates (through five cycles) by training 100 images (until 100% accuracy is achieved) and testing on all 500 images. For training and testing pairs consisting of corresponding 'fa' and 'fb' images the average (cross validated) results obtained using ERBF were as follows:

Pair	False neg. (-)	False pos. (+)
fa & fa	1.6%	0%
fa & fb	32.6%	0.9%
fb & fb	2.8%	0%
fb & fa	29.4%	1.15%

The above results indicate that the probability for false positives is almost nil but that there is a high (about 30%) probability of missing subjects whose pictures were taken at different times (corresponding to the second < fa & fb > and third < fb & fa > scenarios, respectively).

5.3 Discrete Approach

As an appropriate GUI for the manual location and collection of facial features is available we carried out several benchmark studies using the feature (discrete) and correlation (FCA) approach where face identity is determined based on the geometry of facial features using the affine transformations with respect to small rotation and scale - translation was ignored in this case as the origin of (the complex plane) was chosen to be at a fixed feature (the center of the nose). Training and test sets consists of 75 corresponding images drawn from frontals 'fa' and 'fb' each described in terms of sets of V4 (four) or V8 (eight) discrete features. Using cross-validation testing average correct identification ('match') is less than 50%. The false positive error rate is very high while the false negative error rate is almost nil. This last experiment is consistent with earlier experiments using the discrete approach and cast great doubts about the feasibility of using such an approach.

6. Parallel and Distributed Computation

Face recognition performance should eventually involve efficiency as well because both the MATCH and SURVEILLANCE tasks could possibly involve hundred if not thousands of images. The models benchmarked and discussed earlier in this paper, especially the image representation schemes, lend themselves to parallel implementation. To further study computational aspects and the inherently parallel nature of the models referred to in Sections. 5. 1 and 5.2, we have implemented the corresponding dynamic attention schemes on a 64 processor Intel Paragon XP/S (Takacs, Wegman, and Wechsler, 1994). Taking advantage of the fully parallel MIMD environment, a manager node / processor was used to balance the workload and to assign sub tasks to multi purpose worker units, which in turn submit the computed results asynchronously for further investigation. Execution time improved quickly but leveled at its minimum when the number of neurons mapped on a single processor fell within the range of 50 to 75. The use of additional processors (25 and above) did not yield significant speedups causing a steep decrease in efficiency. Besides architectural constraints of the Paragon XP/S - increasing communication overhead and load balancing problems - the results are consistent with psychological and human benchmark studies. Specifically, while adaptation and local feature extraction are parallel in nature, they are followed by

essentially sequential processes required to integrate the information acquired from overlapping and localized receptive fields.

7. Conclusions

The research carried out under the FERET R&D program and the experimental data reported in this paper seem to indicate that (i) future advances in automated face recognition require the development of hybrid recognition systems involving appropriate representation and classification schemes, (ii) that holistic (connectionist) methods outperform discrete (and direct) feature and correlation methods, (iii) that if the test beds are random and contextual cues are lacking human performance is quite poor, not consistent, and degrades rapidly when compared with machine performance, and (iv) that extensive and proper testing is crucial for bench marking. Finally, parallel and distributed computational schemes are essential if performance speedup is required.

References

[1] Samal, A. and P. A. Iyengar (1992), *Automatic recognition and analysis of human faces and facial expressions: A survey*, Pattern Recognition, 25(1), 65-77.

[2] Takács, B. and H. Wechsler (1994), *Locating Facial Features Using SOFM*, 12 Int. Conference on Pattern Recognition, Jerusalem, Israel.

[3] Takács, B. and H. Wechsler (1994), *Face Location Using A Dynamic Model of Retinal Feature Extraction*, Int. Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland (under review).

[4] Takács. B, E. J. Wegman, and H. Wechsler (1994), *Parallel Simulation of an Active Vision Model*, Intel Supercomputer Users Group (ISUG) Conf., San Diego, California.