Human Face Recognition: Problems, Progress and Prospects

1 Introduction

The goal of this report is to relate the extensive face recognition technology which is available in the literature to law enforcement applications. A table of possible applications developed by the FBI is shown in table 1. These applications range form static matching of controled format photographs in mugshot matching, application 1, to real-time matching of surveillence vidio images, application 3. The applications have different advantages and disadvantages and present a wide range of different technical challanges.

In addition to the seperation of images into static and real-time several other parameters are important in evaluating these applications. In any pattern recognition problem the accuracy the solution will be strongly effected by the limitations places on the problem. It is impractal to match a blury photo against a database of photograths sent in bythe entire US population. Two restrict the problem to practical proportions both the image input and the size of the search sapace must have some limits. The limits on the image might for example include controled format, backgrounds which simplify segmentation, and controls on image quality. The limits on the database size might include geographic limits and descriptor based limits. One of the objective of this report is to investigate existing algorithms and ask if the restrictions allowed by the applications present in table 1 make it likely that these algorithms can be used to solve practical law enforcement problems.

Three different kinds problems are presented in table 1; these are matching, similarity detection, and transformation. Applications 1,2, and 3 involve matching one face image to another face image. Applications 4, 5, and 6 involve finding or creating a face image which is similar to the humman recollection of a face. Finally, applications 7 and 8 involve generating an image of a face form input data that is useful in other applications by using other information to perform modifications of a face image. Each of these applications imposes increasingly difficult requirements on the recognition process. Matching requires that the candiate matching face image be in some set of face images selected by the system. Similarity detection requires, in addition to matching, that images of faces be found which are similar to a recalled face; this requires that the similarity measure used by the recognition system closely match the similarity maesures used by humans. Transformation applications require that new images created by the system be similar to human recollections of a face.

Motivation

Problem Statement

Neurophysiological bases

Application	Advantages	Disadvantages
1. Mugshot Matching	controled image controled segmentation good quality consistent with IPS/III	no existing database large potential database rare search type
2. Bank/Store Security	high value geographic search limits	uncontroled segmentation low image quatity
3. Crowd Surveillance	high value small file size	uncontroled segmentation low image quality real-time
4. Witness Face Reconstruction	genitic optomization	unknown similarity
5. Electronic Mugshot Book	descriptor search limits genitic optimization	unknown similarity
6. Electronic Linup	descriptor search limits	unknown similarity
7. Reconstruction of Face from Remains	high value	requires physological input
8. Computerized Aging	missing children	requires example input

Table 1: Law Enforcement Usage of Faces.

What has been done

Shortcomings

What should be done

Evaluation of a face recognition system

Organization of the report

2 Problem Statement

Controlled Vs. uncontrolled

Possible Scenarios

Driver's license, passport pictures

Mug-shots

Face in a cluttered scene

Video sequence

3 Neurophysiological motivations

Ellis in [?] poses a series of ten questions with human face processing. 1) Are faces unique objects? The uniqueness of the face is a view that has dominated the work in face recognition, but the existence of an unique face processing system comes from three source. A) Faces are more easily remembered when presented in an upright orientation than other objects. B) Prosopagnosia patients are unable to recognize previously familiar faces, but usually have no other profound agnosia. C) It is argued that infants come into the world prewired to attracted by faces. Neonates seem to prefer to look at moving stimulus with a face-like patterns in comparison to those containing no pattern or with jumbled features. 2) Is face perception the result of wholistic or feature analysis? Attempts for establishing which facial features attract the most attention and provide useful means for the discrimination faces have perhaps distracted research away from considering the face as a gestalt where inter-feature space may be just as important. Face recognition could make use of a process akin to that employed by caricaturists who exaggerate unusual features. Attempts experimentally to support the idea have proved entirely unsupportive. 3) How useful are information-processing models of face recognition? Theories or models of face processing are fairly new and most of them are in information-processing terms. Some have provided an excellent framework for analyzing the mechanisms underlying face recognition. The models account for the influence of other cognitive processes at different stages. 4) Are identity and expression analysis separable? Basically evidence comes from two sources, tachistoscopic face perception in normals, and face identification and emotional interpretation among patients suffering various forms of dementia. The notion that identity and expression analysis are dissociable is supported by the various findings. 5) What can we learn from machine face recognition? Ellis feels that is a premature question at present. 6) What lessons may learned from disorders in face recognitions? Patients with prosopagnosic are unable to recognize faces but display appropriate autonomic responses to faces. This may indicate the conscious route to identification may not be only means of registering familiarity with a face. 7) Can one improve someone's face recognition ability? The skills for face recognition appear to be such that efforts to improve them are doomed to failure. 8) How does a face become familiar? Familiarity concerns the procedure by which a novel face becomes familiar and once having become familiar, how its representation in memory is updated to account for the changes that inevitably occur over time. 9) How is face memory organized? Direct evidence suggests that face memory organization may be in the form of clustering by physical similarity, interaction of profession, and appearance. 10) How useful are theories of face recognition? As understanding of the above question increases, it is hoped that methods will

be developed to increases witnesses' recall [?].

The study done by Goldstein and Mackenberg in [?] shows that face perception is related to age by using kindergartener. first and fifth graders. The first and fifth grade child is able to recognize more face than a kindergarten child. The study further shows that the older child is able to recognize more faces from partial photographs than an younger child. The recognition does not appear to be a function of familiarity but of age. The study suggests that the upper portions of the face are more helpful in identification than the lower portions [?].

In [?]. Hochberg and Galper discuss the recognition of human faces. It was found that there is a significant difference in recognition accuracy between upright and inverted faces by humans. As a result it is felt that something other than pattern storage and pattern recognition is involved in facial recognition in humans [?].

In [?]. robust methods are protected against departures from the assumptions of analysis. They are also protected against potentially strong influences of atypical or incorrect data values. By allowing regions with large deformations to have a large impact on the fit. least squares methods can minimize true shape differences and obscure them. The resistant fit uses the repeated median algorithm to produce alternatives to the least squares transformation values. The purpose of a fitting method is to choose a frame of reference or comparison whereby the two shapes are superimposed for the observation similarities and differences. If change has taken place in a localized region. the most useful fit is obtained by perfectly superimposing the unchanged region. The changed region can then be easily identified by its lack of fit and large residuals. This desired matching property corresponds to the statistical properties of robustness and resistance. The outlying points that are not consistent with the should not influence the fit. Very different fits can be produced by the least squares and resistant methods. In fitting primate skulls. the leasts squares method and resistant method suggest different allometric relationships. The limits of infulence of the cranial vault area using the resistant fit achieves a closer agreement in the facial region than the least squares method. Several large residual vectors stand out against the background of many small ones. The resistant fit provides a clear basis for the identification of differences and similarities of shape. Residual vectors from a least squares fit generally tend to cluster near an average length blurring the distinctions and rendering inferences more difficult. Least squares methods have proven themselves in many situations. Least squares methods produce an overall fit whose residual sum of squares is a useful single-number measure of how different two specimens are. In analyzing comparative morphology. the interest is in the detailed identification of similarities and allometric shape differences. than a resistant method. such as the repeated median algorithm would be preferable [?].

4 Face Recognition from Still Intensity and Range Images

4.1 Segmentation

Craw et al. in [?] describe a method for extracting the head area from the image. They proceed by using a hierarchical image scale and template scale. Constraints are imposed on the location of the head in the image. Resolutions of $8 \times 8, 16 \times 16, 32 \times 32, 64 \times 64$ and full scale 128×128 are used in their multiresolution scheme. At the lowest resolution a template is constructed of the head outline. Edge magnitude and direction are calculated from the gray level image using the Sobel mask. A line follower is used to connect the outline of the head. Since connectivity is based on similar edge direction and magnitude, it can be easily confused with all possible edges present in the image which are other than the head outline. To compensate for such a condition, the template is used as a guideline to find the head outline. Here it becomes apparent why staggered image resolutions are used. The result of lower resolution are used as the guidelines for the next higher resolution, since the chances of error are at their minimum for the lowest resolution. After the head outline has been located the search for lower level features such as eyes, eyebrows, and lips is conducted according to pattern search, guided by the location of the head outline, using a similar line following method. The results depicted show a reasonably good chance of detecting the head outline though the search for the eyes location did not fare well. An improvement in the system could be achieved by using the Canny's [?] or Burr's [?] edge finder. However, in a later work by the author, this technique was replaced by a wire mesh method.

In [?] Craw, Tock and Bennet describe a system to recognize and measure facial features. Their work was motivated in part by automated indexing police mugshots. They endeavor to locate 40 feature points from a grey-scale image, these feature points were chosen according to Shepherd [?] which was also used as a criteria of judgment. The system uses a hierarchical search coarse to finer. The template drew upon the principle of polygonal random transformation in Grenander et al. [?] and [?]. The approximate location, scale and orientation of the head is obtained by iterative deformation of the whole template by random scaling, translation and rotation. A constraint of feasibility is imposed so that these transformation do not end up having no resemblance to the human head. Optimization is achieved by simulated annealing. After a rough idea of the location of the head is obtained, a refinement is done by transforming individual vectors of the polygon, details of which are in [?]. The authors proclaim successful segmentation of the head in all 50 images that were tested. In 43 of these images a complete outline of the head was distinguishable, the remaing

faulted at finding the chin. The detailed template of the face included eyes, nose, mouth etc., in all 1462 possible feature points were searched for. Feature experts (templates) were used for this stage. The authors claim to be able to identify 1292 of these feature points. The only missing feature was that of the eyebrow as they did not have a feature expert for that. They attribute the 6% incorrect identification to be due to beards and moustaches present in their database which caused mistakes in locating the chin and the mouth of the subject. It should be noted that due to the optimization and random transformation, the system is inherently computationally intensive. The authors mention further work related to obtaining the eye blink rate from a sequence of images.

Govindaraju et al. [?] consider a computational model for locating the face in a cluttered image. Their technique utilizes a deformable template which is slightly different than that of Yuille et al. [?]. Working on the edge image they base their template on the outline of the head. The template is composed of three segments that are obtained from the curvature discontinuities of the head outline. These three segments form the right side-line, the left side-line and the hairline of the head. Each one of these curves is assigned a four tuple comprising of the length of the curve, the chord in vector form, the area enclosed between the curve and the chord, and the centroid of this area. To determine the presence of the head, all three of these segments should be present in a particular orientation. The center of these three segments gives the location of the center of the face. The templates are allowed to translate, scale and rotate according to certain spring based models. They construct a cost function to determine hypothesized candidates. Details are present in [?]. They have experimented on about ten images and though they claim to have never failed to miss a face, they do get false alarms.

4.2 Feature Extraction

One of the earlier works on face recognition was done by Sakai et al. in [?]. They used a digitized image with eight grey levels. The work was conducted on a data set comprising of frontal face images. A 3×3 mesh was used to determine pixels of the greatest gradient values so that the amount of information is reduced to the bare essential. These pixels were then connected to neighboring pixels exhibiting similar characteristics to form line and contour segments. Details are present in [?]. The recognition principle used was pattern matching using a pre-defined face template. A course to fine approach is used to determine individual features of the face. It should be noted that the recognition does not differenciate amongst different faces rather, it determines the existance of a face in the image. The authors note that the procedure they employ has dependency on the illumination direction. Changes

made to the illumination direction cause problems to their approach. The same has been commented on by various other researchers as well and poses to be a consistant problem.

Reisfeld and Yeshurun in [?] describe a generalized symmetry operator for the purpose of finding the eyes and mouth of the face. Their motivation stems from the almost symmetric nature of the face about a vertical line through the nose. Subsequent symmetries lie in the features such as eyes, nose and the mouth. The symmetry operator locates points in the image corresponding to a high symmetry measure discussed in detail in [?] . From their paper it is not apparent how they determine exactly which feature is the eye or the nose i.e. no qualifiers are given for this determination. They have mentioned their procedures superiority over other correlation based schemes like Baron [?] in the sense that their scheme is independent of any scale or orientation dependencies. However, since no a priori knowledge of face location is considered, the search for the symmetry points will be computationally intensive. The authors mention success rate of 95% on their face image database, with the constraint that the face occupy between 15-60% of the image.

Yuille, Cohen and Hallinen in [?] extract facial features using deformable templates. These templates are allowed to translate, rotate and form to fit the best representation of their shape present in the image. Certain preprocessing is done to the initial intensity image to get representations of peaks and valleys from the intensity image. Morphological filters are used to determine these representations. They constructed a template for the eye with eleven parameters comprising of the upper and lower arc of the eye: the circle for the iris; the center points; and the angle of inclination of the eye. The template is fit to the image in the energy minimization sence. Energy fuctions of valley potentials, edge potential, image potential, peak potential and internal potential are determined. Coefficients are selected for each potential and an update rule is employed to determine the best parameter set. Details are in [?]. In their experiments they found that the starting point of the template is criticle for determining the exact location of the eye. When the template was started above the eyebrow, the algorithm failed to distinguish between the eye and the eyebrow. Another draw back to this approach is its computational complexity, it takes 5 to 10 minutes on a SUN 4 with a good starting point selected. Generally speaking template based approaches to feature extraction are a more logical approach to take. The problem lies in the description of these templates. Whenever analytical approximations are made to the image the system has to be tolerant to certain discrepencies between the template and the actual image. This tolerence tends to average out the differences that make individual faces unique.

Gillenson and Chandrasekaran [?] describe a computerized male facial compositor. They build upon the assumption that an artist's sketch of the face has to be able to incorporate distinct facial features to be a good rendering of the original face image. As the artist starts

composing a sketch, he/she has to develop an outline then build upon that. Their man-machine system relies on human inputs to visual/verbal cues to construct a sketch of the face. The system has in its database 17 features of the face with seven of them as pairs. These are: hair outline, chin, eye orbits including pupils (paired), upper and lower eyelids (paired) including crows feet, eyebrows (paired), forehead lines, upper and lower part of the nose, upper and lower lip including center mouth line, auxiliary chin lines, cheek lines (paired), naso-labial lines (paired), ears paired and neck lines. These features are present in all the different possible forms that will allow them to be fit to the face in question with scaling, translational and rotational cues provided by the user. A user of the system starts working with the sketch of an average face. The sketch is aged according to user input. Changes to the sketch occur hierarchically from the head outline to the internal features like eyes, nose etc. Shading is done as the last step. Experimental results show a marked improvement between hand drawn and computer aided sketches by nonartist people when judged by a jury of 62 persons. On the whole this system was not developed with the aim of automated face recognition. However, a description of face into features that are unique to individuals either one by one or a collection thereof provides a basis for any automated face recognition system. To be able to quantify the information that the artist uses in sketching a face should improve the recognition capability of a face identification system.

In [?], the image features are divided into four groups, which are: visual features, statistical features of the pixel, transform co-efficient features, and algebraic features. Hong concentrates on the algebraic features which represent the intrinsic attributes of an image. The Singular Value Decomposition (SVD) of a matrix is used to extract the features from the pattern. The singular values (SV) extracted by the SVD have a good performance as a shape descriptor. In Hong's view the SVs of an image are very stable and represent the algebraic attributes of the image being intrinsic but not necessarily visible. [?] proves the stability and invariances to proportional variance of image intensity in the optimal descriment vector space, to a transposition transform, rotation transform, to translation, and mirror transform with are important properties of the SV feature vector. Hong states that an image $A_{m\times n}$ is represented as an n-dimensional SV feature vector. The image recognition problem may be solved in an n-dimensional feature space. A facial photo of 32mm $\times$ 27mm would need a 70-dimensional SV feature vector to describe it. The original high dimension SV feature vector may be compressed low dimension feature space (2-D or 1-D) using various transforms. The Foley-Sammon transform is used to obtain the optimal set of discriminant vectors spanning the Sammon discriminant plane. With a small set of photos, Hong uses only two of the vectors for recognition but predicts that more of the discriminant vectors will be needed for recognition with more photos. The small set of photos consists of 9 facial photos of 50mm $\times$

35mm. Each photo was sampled five times by varying the relative position between photo and the TV camera used for scanning for a total of 45 image samples with 5 images in each class. The SVD operation was applied to each image matrix for the extracting of SV features and the SV vector is obtained. The optimal discriminant plane and quadratic classifier of the normal pattern is constructed for the 45 SV feature vector samples. The classifier is able to recognize the 45 training samples of the 9 subjects. Testing is done using 13 photos which consists of 9 newly sampled photos of the original test subjects with two of one subject and 3 samples of different ages of the subjects involved in the training set. There is a 42.67% error rate which Hong feels are due to the statistical limitations of the small number of training samples [?].

Cheng in [?] states that classification approaches based on structure parameters are not robust for recognizing complex human face images. Structure parameters are not stable enough. They sensitive to the changes in rotation, scaling, and facial expressions. Hong in [?] suggest the use of the singular value (SV) vectors, which are sensitive to geometric variance, such as rotation and scaling. The development of shape descriptors [?], the shape of an object is independent described regardless of any translation or rotation. Cheng's $A_{m\times n}$ image in [?] is represented by an n-dimensional SV vector. The SV vector is compressed into a low dimensional space by means of various transforms, the most popular being an optimal discriminant transform based on Fisher's criterion. The Fisher optimal discriminant vector represents the projection of the set of samples on the direction φ. It ensures that the patterns have a minimal scatter within each class and a maximal scatter between class in the 1-dimensional space. Three SV feature vectors are extracted from the training set in [?]. The optimal discriminant transform compresses the high dimensional SV feature space to a new $r-$dimensional feature space. The new secondary features are algebraically independent and informational redundancy is reduced. Cheng's experiment is performed on 64 facial images from 8 people (classes). The photographs are represented by Goshtasby's shape matrices, which are invariant to translation, rotation, and scaling of the facial images in the form the matrix and is obtained by a polar quantization of a shape [?]. Three photographs from each class are used as a training set of 24 SV feature vectors. The SV feature vectors are treated with the optimal discriminant transform to obtain the new second feature vectors for the 24 training samples. The class center vectors are obtained using the second feature vectors. The experiment uses six optimal discriminant vectors. The separability of training set samples is good and there is 100% recognition of the training set. The remaining 40 facial images are used the test set, 5 from each person. Changes are made in the relative camera position and the face, the camera's focus, the camera's aperture setting, the wearing or not wearing of glasses, and blurring. As with training set, the SV feature vectors are extracted.

the optimal discriminant transform is applied obtaining the second feature vector. Again good separability is obtained and there is an accuracy rate of 100% [?].

4.3 Recognition

In [?] a basic study in classifying human faces is reported. The photographs used are the front view of human faces, with the mouth closed, no beards, and no eyeglasses. The individuals are looking into the camera when the photographs are taken. To Kaya et al. the parameters for characterizing a face need to be easily estimated from the photographs with changes in lighting and degrees of development having little effect, small changes of facial expression having little effect, and carrying as much information as possible. Kaya et al. estimates that the number of parameters should greater than $\log_2 N$ bits where N equals the number of faces to be classified. In [?] the use of a set Fourier coefficients is rejected as a parameter as the intensity of the photograph and its distribution depend on the lighting conditions and development technique. Kaya et al. uses Euclidean distances between singular points on the face as the characteristic parameters. The characteristic parameters are normalized by dividing by the nose length to account for any differences due to the size of the photograph and the distance the subject is from the camera. The photographs of 62 Japanese adults between the ages of 20 and 30 are uses. The photographs have been taken under the same lighting conditions. The characteristic parameters are measured by hand. The mean and standard deviation are calculated for the parameters. The parameters' correlation indicates that the actual dimension of the parameter vector may be smaller 9, the number of parameters. Equal weighting of the parameters is a result of normalization. The parameter matrix is dominated by the principal eigenvector as a result of the normalization. The components of the principal eigenvector in the original parameter space are positive. The number of classifiable faces depends largely on the algorithm of classification. One metric for the effectiveness of the classification algorithm is the average number of parameters used to identify the face. A smaller the number of parameters better as it decreases the needed amount of storage the the patterns. The workability of algorithm is based in the probability of finding the correct face and the expected number steps needed until the identity is found. For the 9 geometric parameters used by [?], it has been found that the amount of stored information may range from 8 to 14 bits depending of the noise estimation. The noise components are: the inherent noise D_i and the noise present in extracting the parameters (D_m) from a given photograph. D_m consists of noise in detecting ambiguous facial parts' outlines and quantization noise from the conversion of a photograph into the small picture elements of which densities are inputs to the computer. D_m's properties depend on the camera's characteristics and algorithm for extracting the parameters from the photographs.

D_i is the variation of parameters of a face from photograph to photograph. d_i is caused by variations in rotation, expression, etc. Variations in the geometry of face over a short period of time are considered to be D_i. The amount of information carried by the parameters is average of mutual information of the parameters and the estimated total noise [?].

The automation of the identification of human faces consists of three problems: the selection of effective features to identify the faces, machine extraction of those features, and decision-making based on the feature measurements. One method to characterize the face is the use of geometrical parametrization i.e. distances and angles among the points such as eye corners, mouth extremities, nostrils and chin top [?]. The data set used by Kanade consists of 17 male and 3 female faces without glasses, mustaches, or beards. Two pictures have been taken of each individual, with the second picture being taken one month later in a different setting. The face-feature points are located in two stages. The coarse grain stage eases the succeeding differential operation and feature-finding algorithms. Once the eyes, nose and mouth are approximately located, more accurate information is extracted by confining the processing to four smaller regions and scanning at higher resolutions and using the "best beam intensity" for the region. The four regions are the left and right eye, nose, and mouth. The beam intensity is based on the local area histogram obtained in the coarse grain stage. Kanade regards the rescanning process as a type of visual accommodation in both light intensity and location. After the finer image data acquisition, more precise information is extracted from each region using thresholding, differentiation, and integral projection. The two stage process may reduce memory and time. Kanade uses a set of sixteen facial parameters which are ratios of distance, area, and angels to compensate for the varying size of the pictures. To eliminate scale and dimension differences the components of the resulting vector are normalized. The entire data set of 40 images is processed and one picture of each individual is used in the reference set. The remaining 20 pictures are used as a test set. Kanade uses a simple distance to measure the similarity an image of the test set and the image in reference set. Kanade's results range from 45% to 75% depending on the parameters used. When several of the ineffective parameters are not used the better results are obtained [?].

[?] deals with the use of a computer in the storage and retrieval of mug-shots. Such a system has two steps in pattern recognition, feature selection and the algorithm. Three distinctions are made in the prototype development efforts, the nature of the target image that serves as input to the search, the method of coding the faces, and the pose position of the faces in the mug-file. [?] presents a review of several systems developed to cope with what type of input to use in recognizing and retrieving a face from the gallery. [?] reviews three prototype systems with varying methods of interaction and input. "Whatsisface", "Sketch",

and "computerized montage" systems. Laughery et al. also review the different interactive and automatic measurement algorithms for locating and measuring facial features. With the interactive system, facial images can be measured accurately with a precision which exceeds manual methods. Automatic measurement systems are not completely reliable, yet. The basic approach to the mug-file problem is for the system to compare features from the target with those stored in the data base. The nature of the target image relative to the images in the data base is crucial and determines the difficulty of the overall procedure. The target may be a mug-shot or from another photographic source and need to be rotated before the features can be extracted and compared to the mug-file images. In the case of line drawings, the quality of the image depends on the artist's ability and talent. The features selected for use in pattern recognition system need to be discriminatory, easily obtained, stable over time, efficient and economic during data collection, storage and application. The precision and accuracy of measuring and encoding a feature are important in terms of the useful information provided and the cost of using the feature. Precise and accurate measurement of a feature with less information content may be more cost-effective than imprecise measurement of with more information content. Once a set of features has been measured for a larger set of faces, statistical analyses can be performed to identify a subset of features which t end to be independent and summarize the information in the original set. The number of features is an important parameter in a pattern recognition system. Features may be described syntactically or geometrically. A geometric system of coding usually combines basic measurements into features summarizing the image information. In [?] 10 feature distances are measured to code 9 features with each feature being a distance divided by the nose length. They found indications that 92% of the variation in the normalized data could be explained by 5 components. The components are considered high-level features when they summarize the information from the basic features. Similarity measures are used in sequencing algorithms with geometric coding of features. The objective of this approach is to sequence the photographs in the mug-file on the basis of similarity to the target. The algorithm's design must consider the precision and accuracy of the measurements and develop appropriate scales for the components of the feature vector. The matching algorithm can use either syntactic data or geometric data. The computer is programmed to select matching features in a specified order. Algorithm designing is concerned with the sequence in which selections are made and how to minimize errors in the face description. In [?]'s matching algorithm uses a window for each feature and selects all the images that fall within the window. A larger window permits more of the data base's population to fit, while a smaller window increases the probability that an error in coding a feature will cause the correct image to be missed during the search. Harmon et. al with geometric feature codes of images, used a matching algorithm to elim-

inate the mismatches and a sequencing algorithm on the remaining subset achieved 'nearly perfect' identification for a population of over 100 faces. It is felt that either the matching or sequencing algorithms may be used to recognize and retrieve facial images [?].

In [?] Haig uses an image-processing computer to increase the understanding of face recognition. The system advantages include the abilities to insert new targets into digital storage, to control and change the intensities both locally and globally, to move the targets around, to change their size and orientation, to present them for a wide range of fixed time intervals, to run experiments automatically, to collect the data, and to analysis the results. Haig's database consists of 100 target faces, taken under reasonably standardized conditions from the direct frontal aspect. The pictures are stored in 128×128 pixels, and registrered such that the inter-pupilary distance is 30 pixels. The goal of Haig's face distortion experiments is to measure the sensitivity of adult observers to slight positional changes of prominent facial features. Each target face is subjected to the same operations, in which certain features are moved by defined amounts. Greatest sensitivity in movement is to the movement of the mouth up at 1.2 pixels, close to the visual acuity limit. In other experiments, features are interchange among four faces different faces. The head outline is the major focus of attention when four features are interchanged. The observers scored 28.7% correct for the head, 24.3% for the eyes and mouth, and 22.7% for the nose. The pattern of correct responses as a function of the changed single-feature, heads scored 34.0%, mouths scored 8.9% , eyes scored 8.3%, and noses scored 0.5%. A changed head outline, while maintaining the inner features very strongly influences the observers responses. Fixing the head position, Haig tests the inner features. The results show that the observers favor the eyes with 58.7%, mouths with 24.0% and noses with 17.3%. The response pattern demonstrates a most gratifying consistency: Response to changed eyes is 80.3%, changed mouths is 13.8% and changed nose is 1.3% . Fixing both the head and the eyes shows a dominance in the mouth over the nose. These experiments establish a clear hierarchy of feature saliency in which the head's outline plays a major role. In the distributed aperture experiments, Haig attempts to find what constitutes a facial recognition feature. The technique used implies that all parts of the face are equally likely to be masked or unmasked in any combination. The program selects one of the four target faces at random and presents the target to a random number of apertures with their actual addresses selected at random from the 38 possible addresses. Haig in looking at the overall results, noticed the very high proportion of correct responses across the eyes-eyebrows and across the hairline at the forehead. Few correct responses may be seen around the side of the temples and at the mouth, and the lower chin area is clearly not a strong recognition feature. In the higher resolution distributed apertures experiment, it is determined that the resolution could be doubled, in each orthogonal direction. This

allows for a total of 162 usable apertures on each target. The median face area required for recognition in the lower aperture experiment is 5.9% and for the higher aperture experiment requires 4.0%. The reason for the lower threshold for the experiment is the greater variety of aperture positions and combinations [?].

Recently, the use of Karhunen-Loeve (KL) expansion for the representation [?] and recognition [?] of faces has generated renewed interest. The KL expansion has been studied for image compression for more than twenty years [?]; its use in pattern recognition applications has also been documented for quite some time [?]. One of the reasons why KL methods, although optimal did not find favor with image compression researchers is their computational complexity. As a result, fast transforms such as the discrete sine and cosine transform have been used [?]. In [?], Sirovich and Kirby revisit the problem of KL representation of images (cropped faces). Nothing that the number of images M usually available for computing the covariance matrix of the data is much less than the row or column dimensionality of the covariance matrix, leading to the singularity of the matrix, they use a standard method from linear algebra [?] that will calculate only the M eigenvectors that do not belong to the null space of the degenerate matrix. Once the eigenvectors (referred to as eigenpictures) are obtained, any image in the ensemble can be reconstructed using a weighted combination of eigenpictures. By using increasing number of eigenpictures, one gets improved approximation to the given image. The authors also give examples of approximating an arbitrary image (not included in the calculation of eigenvectors) by the eigenpictures. The emphasis in this paper was on the representation of human faces.

In a subsequent extension of their work, Kirby and Sirovich in [?] include the inherent symmetrization in faces in the eigenpicture representation of faces, by using an extended ensemble of images made of original faces and their mirror symmetries. Since the eigen computations can be split into even and odd pictures, there is no overall increase in the computational complexity compared to the case when only the original set of pictures in used. Although the eigenrepresentation for the extended ensemble does not produce dramatic reduction in the error in reconstruction when compared to the unextended ensemble, still the method that accounts for symmetry in the patterns is preferable.

Turk and Pentland [?] used the eigenpictures (also known as eigenfaces in [?]) for face detection and identification. Given the eigenfaces, every face in the database can be represented as a vector of weight; the weights are obtained by projecting the image into eigenface components by a simple inner product operation. When a new test image whose identification is required is given, the new image is also represented by its vector of weights. The identification of the test image is done by locating the image in the database whose weights are the closest (in Euclidean distance) to the weights of the test image. By using

the observation that the projection of a face image and a non-face image are quite different, a method for detecting the presence of a face in a given image is given. Turk and Pentland illustrate their method using a large database of 2500 face images of sixteen subjects, digitized at all combinations of three head orientations, three head sizes and three lighting conditions. Several experiments were conducted to test the robustness of the approach to variations in lighting, size, head orientation and the differences between training and test conditions. Impressive recognition rates are reported for a reasonable size database of 2500 images. It is reported that the approach is fairly robust to changes in lighting conditions, but degrades quickly as the scale changes. One can explain this by the significant correlation present between images with changes in illumination conditions; the correlation between face images at different scales is rather low. Another way to interpret this is that, the eigenfaces approach will work well as long as the test image is "similar" to the ensemble of images used in the calculation of eigenfaces. Turk and Pentland, also extend their approach to real time recognition of a moving face image in a video sequence. A spatio-temporal filtering step followed by nonlinear operation is used to identify a moving person. The head portion is then identified using a simple set of rules and handed over to the face recognition module.

In [?], the Karhunen-Loeve (KL) transform is used for the extraction of features from face images. The KL is combined with a two other operations to improve the performance of the extraction technique for the classification of front view faces. The application of the KL expansion directly to a facial image without standardization does not achieve a robustness against the variations found in image's acquisition. [?] uses a standardization of the position and size of the face. The center points are the regions corresponding to the eyes and mouth. Each target image is translated, scaled and rotated through affine transformation so the reference points of the eyes and mouth satisfy a specific spatial arrangement with a constant distance. An empirically defined standard window encloses the transformed image. The KL expansion is applied to the standardized face images and is known as Karhunen-Loeve transform on Intensity Pattern in the Affine transformed Target image feature (KL-IPAT). The KL-IPAT is extracted from 269 images with 100 eigenfaces. The KL-IPAT greatly improves image recognition when compared with the eigenface approach using the KL on the raw image. The second step is to apply the Fourier Transform to the standardized image and use the resulting Fourier spectrum instead of the spatial domain of the standardized image. The KL expansion is applied the Fourier spectrum and the extracted feature is called the Karhunen-Loeve transform of Fourier Spectrum in the Affine transformed Target image feature (KL-FSAT) and uses of the same 269 images. The robustness of the KL-IPAT and KL-FSAT is checked against geometrical variations using the standard feature for 269 individuals. In the first experiment, the training and testing samples are acquired in as

similar conditions as possible. The test set consists of 5 samples from 20 individuals. The KL-IPAT has an accuracy rate of 85% and the KL-FSAT has an accuracy rate of 91%. Both methods miss identified the one example where there is a difference in the wearing and not wearing of glasses between the testing set and the training set. The second experiment checks for feature robustness when there is a variation caused by an error in the positioning of the target window. This is an error usually made during image acquisition due to changing conditions. The test images are created by shifting the reference points various directions by one pixel. The variances for 4 and 8 pixels are tested. The KL-IPAT having an error rate of 24% for the 4 pixel difference and 81% for the 8 pixel difference. The KL-FSAT has an 4for the 8 pixel difference. The improvement is due to the shift invariance property in the Fourier spectrum domain. The third experiment uses the variations in head positioning. The test samples are taken while the subject is nodding and shaking his head. The KL-FSAT shows a high robustness over the KL-IPAT in the different orientations of the head. Good recognition performance is achieved by restricting the image acquisition parameters. Both the KL-IPAT and KL-FSAT have difficulties when the head orientation is varied [?].

The use of isodensity lines, boundaries of constant gray levels for face recognition has been investigated in [?]. Isodensity lines, although are not directly related to 3-D structure of a face do represent a relief of the face. Using images of faces taken with a black background, a Sobel operator and some post processing steps are used to obtain the boundary of the face region. The gray level histogram (an 8 bin histogram) is then used to trace contour lines on isodensity levels. A template matching procedure is used for face recognition. The method has been illustrated using ten pairs of face images, with three pairs of pictures of men with spectacles, two pairs of pictures of men with thin beard and two pairs of pictures of women. Good recognition accuracy is reported on this small data set.

The use of Neural Networks (NN) in face recognition has addressed several problems: gender classification, face recognition and classification of facial expressions. In [?], Golomb, Lawrence and Sejnowski present a cascade of two neural networks for gender classification. The first stage is an image compression NN whose hidden nodes serve as inputs to the second NN that performs gender classification. Both networks are fully connected, three-layer networks with two biases and are trained by a standard back-propagation algorithm [?]. The images used for testing and training were acquired such that facial, hair, jewelry and makeup were not present. They were then preprocssded so that the eyes are level and eyes and mouth are positioned similarly. A 30X30 cropped block of pixels was extracted for training and testing. The dataset consist of 45 males and 45 females; 80 were used for training, with 10 serving as testing examples. The compression network indirectly serves as a feature extracter; in that the activities of 40 hidden nodes (in a 900 X 40 X 900)

networks serve as features for the second network the performs gender classification. The hope is that due to the nonlinearities in the network, that feature extraction step may be more efficient than the linear K-L methods [?]. The gender classification network is a 40 X n X 1 network, where the number of hidden nodes has been one of the followings 2,5,10 20, or 40. Experiments with 80 training images and 10 testing images have shown the feasibility of this approach. This method has also been extended to classifying facial expressions into one of eight types.

Using a vector of sixteen numerical attributes such as eyebrow thickness, widths of nose and mouth, six chin radii etc. Brunelli and Poggio [?] also develop a NN approach for gender classification. They train two HyperBF networks [?], one for each gender. The input images are normalized with respect to scale and rotation by using the positions of eyes detected automatically. The sixteen dimensional feature vector is also automatically extracted. The outputs of the two HyperBF net works are compared, the gender label for the test image being decided by the network with greater output. For actual classification experiments only a subset of the 16 dimensional feature vector is used. The database consists of 21 males and 21 females. The leave-one-out strategy [?] was employed for classification. When the feature vector from the training set was used as the test vector, 92.5% correct recognition accuracy is reported; for faces not in the training set, the acuracy further dropped to 87.5%. Some validation of the automatic classification results has been reported by using humans.

By using an expanded 35 dimensional feature vector, and one HyperBF per person, the gender classification approach has been extended to face recognition. The motivation for the underlying structure is the concept of a grandmother neuron: a single neuron (the Gaussian function in HyperBF network) for each person. As there were relatively few training images person, a synthetic data base was generated by per perturbing arround the average of feature vectors of available persons and the available persons were used as testing samples. For different sets of tuning paremeters (coefficients, centers and metrics of the HyperBf's) classification results are reported. Some corraboration of the caricatural behavior of the HyperBF networks with psychophysics studies is also presented.

The use of HyperBF networks for face recognition is reported in [?]. To remove variations due to changing viewpoints, the images are first transformed using 2-D affine transforms. The transformation parameters are obtained by using the detected positions of eyes and mouth in the given image and the desired posit ions of these features. The transformed image is then subjected to a directional derivative operator to reduce the effects of illumination source. The resulting image is multiplied by a Gaussian function and integrated over the receptive field to achieve dimensionality reduction. The MIT media lab database of 27 images --person, of 16 different persons is used with images of 17/ persons being used for training, while the rest

being used as testing samples. A HyperBF is trained for each person. Reasonable results are reported. By feeding the outputs of sixteen HyperBF to another HyperBF, significant reductions is error rates are reported.

[?] presents the results of work using a connectionist model of facial expression. The model uses the pyramid structure to represent image data. Each level of the pyramid is represented by a network consisting of one input, one hidden, and one output layer. The input layers of the middle levels of the pyramid are the outputs of the previous level's hidden units when training is complete. Network training at the lowest level is carried out conventionally. Each network is trained using a fast variation of the back propagation learning algorithm. The training pattern set for the subsequent levels is obtained from the combining and partitioning the hidden units' outputs of the preceding level. The original images of the training set are partitioned into blocks of overlapping squares. The overlapping blocks are to simulate the local receptive fields of the human visual system. Each block consists of the set of block patterns partitioned in the same position over the image patterns set. The data set for training consists of 6 hand drawn faces with 6 different expressions: happy, surprise, sad anger, fear and normal. The outer features of each face are shape and ears. The inner features are the eyebrows, the eyes, the nose, and the mouth. Each face is drawn to be as dissimilar as possible from another. The testing set consists of the 6 training faces and the images from the training set masked with a horizontal bar across the upper, middle, and lower portions of the face covering approximately 20% of the total image. The horizontal bar is used to demonstrate the network's associative memory capability. The network has 4 levels. Levels 1-3 consist of 25 input units, 6 hidden units, and 25 output units. The fourth level has 18 input units, 8 hidden units, and 25 output units. The network training process at each level results in a different representation of the original image data. The last level of the pyramid has the leanest and most abstract representation. The representation is viewed as an unique identification of the face and the information it conveys. The network is able to successfully recognize the members of the training set when tested on them. The network poorly recognizes (50%) the various masked, blurred, or distorted facial expressions. It is unable to recognize the various masks of the happy face. The error rate is the result of obtaining a totally different abstract representation which network has not learned. On analysis of the hidden units, patches are found. The patches block off some of the features of the faces and appear unimportant to the hidden node. The hidden units' internal representations show that many of them are in the form of eigenfeatures where the features of the faces are combined in an overlaying manner on top of each other. The eigenfeatures are only a portion of all the features. In the happy face the blocked patches of the hidden units are mainly outside of the face while the others are in

the face. This may be explained in that the happy face does not have many facial features in common with the other faces in the training set. It appears that the network developed a holistic representation of the happy face so that it may be recognized. In conclusion, the leaner representations of the face are automatically generated and are an unique identification of the learned object. The unique representation may be associated with the original object in the form of one-to-many. The model is able to successfully identify the same face but not the masked faces of the same type. The masking of areas shows where the network's learning is focused. It appears that the middle portion of the face image are not as important as the upper and lower and may be used to develop a focus of attention [?].

The systems presented in [?] and [?] are based on the Dynamic Link Architecture (DLA). DLAs attempt to solve some of the conceptual problems of conventional Artificial Neural Networks. The most prominent problem being the expression of syntactical relationships in neural networks. The DLAs use synaptic plasticity and are able to instantly form grouped sets of neurons into structured graphs and maintain the advantages of neural systems. A DLA permits pattern discrimination with the help of an object-independent standard set of feature detectors, automatic generalization over large groups of symmetry operations, and the acquisition of new objects by one-shot learning reducing the time-consuming learning steps. Invariant object recognition is achieved in respect to background, translation, distortion and size by choosing a set of primitive features which is maximally robust with respect to such variations. Both [?] and [?] use Gabor based wavelets for the features. The wavelets are used as feature detectors, characterized by their frequency, position, and orientation. Two nonlinear transforms are used to help during the matching process. A minimum of two levels, the image domain and the model domain, are needed for a DLA. The image domain corresponds to primary visual cortical areas and the model domain to the intertemporal cortex in biological vision. The image domain consists of a 2-D array of nodes $A_x^I = (x, \alpha)$— $\alpha = 1....F$. Each node at position x consists of F different feature detector neurons (x, α) as attributes forming the local descriptors of the image. The label α is used to distinguish different feature types. The amount of feature type excitation is determined for a given node by convoluting the image with a subset of the wavelet functions for that location. Neighboring nodes are connected by links, encoding the information about the local topology. Images are represented as attributed graphs. Attributes attached to the graph's nodes are activity vectors of local feature detectors. An object in the image is represented by a subgraph of the image domain. The model domain is an assemblage of all the attributed graphs, being idealized copies of subgraphs in the image domain. Excitatory connections are between the two domains and are feature preserving. The connection between domains occurs if and only if the features belong to corresponding feature types. The DLA machinery is

based on a data format able to encode information on attributes and links in the image domain and to transport that information to the model domain without sending the image domain position. The structure of the signal is determined by three factors: the input image, random spontaneous excitation of the neurons, and interaction with the cells of the same or neighboring nodes in the image domain. Binding between neurons is encoded in the form of temporal correlations and is induced by the excitatory connections within the image. Four types of bindings relevant to object recognition and representation: binding all the node and cells together that belong to the same object, expressing neighborhood relationships with the image of the object, bundling individual feature cells between features present in different locations, and binding corresponding points in the image graph and model graph to each other. DLA's basic mechanism, in addition to the connection parameter between two neurons, is a dynamic variable (J) between two neurons (i, j). J-variables play the role of synaptic weights for signal transmission. The connection parameters merely act to constrain the J-variables. The connection parameters may be changed slowly by long-term synaptic plasticity. The connection weights J_{ij} are subject to a process of rapid modification. J_{ij} weights are controlled by the signal correlations between the neurons i and j. Negative signal correlations lead to a decrease and positive signal correlations lead to an increase on J_{ij}. In the absence of any correlations, J_{ij} slowly returns to a resting state. Rapid network self-organization is crucial to the DLA. Each stored image is formed by picking a rectangular grid of points as graph nodes. The grid is appropriately positioned over the the image to be stored and is stored with each grid point's locally determined jet and serve as the pattern classes. New image recognition takes place by transforming the image into the grid of jets, and all stored model graphs are tentatively matched to the image. Conformation to the DLA is done by establishing and dynamically modifying links between vertices in the model domain. During the recognition process an object is selected from model domain. A copy of models' graph is positioned in a central position in the image domain. Each vertex in the model graph is connected to the corresponding vertex in the image graph. The matches quality is evaluated using a cost function. The image graph is scaled by a factor while keeping the center fixed. If the total cost is reduced the new value is accepted. This is repeated until the optimum cost is reached. The diffusion and size estimation are repeated for increasing resolution levels and more of the image structure is taken into account. Recognition takes place after the optimal total cost is determined for each for each object. The object with the best match to the image is determined. Identification is a process of elastic graph matching. The process is based predetermined threshold, the image is classified as unknown to the system. In the case of faces, if one face model matches significantly better than all competitor models, the face in the image is considered as recognized. The system identifies

a person's face by comparing an extracted graph with a set of stored graphs. In [?] the experiment consists of a gallery of over 40 different faces images and with little effort to standardize the images, the system recognition success is remarkably consistent. The system shows that a neural system gains power when provided with a mechanism for grouping. The system used in [?] has a larger gallery of faces and recognizes them under different types of distortion and rotation in depth achieving less then 5% false assignments. Lades et. al state that when a clear criterion on the significance for the recognition process is determined, all false assignments are rejected and no image is accepted if its corresponding model is temporarily removed from the gallery. This means that the capacity of the gallery to store distinguishable objects is certainly larger than it present size. No limits to this capacity other than a linear increase in computation time have encountered so far. The system is processing-intensive. Most of the time is spent on image transformation and on optimizing the map between the image and individual stored models [?] [?].

5 Face Recognition from Profiles

Kaufman and Breeding developed a face recognition system using profile silhouettes. The image acquired by a black and white TV camera is thresholded to produce a binary, black and white, the black corresponding to the face region. A preprocessing step then extracts the front portion of the silhouette that bounds the face image. This is to ensure variations in the profile due to changes in hairline. A set of normalized autocorrelations expressed in polar coordinated is used as a feature vector. Normalization and polar representation steps ensure invariance to translation and rotation. A distance weighted k nearest neighbor rule is used for classification. A procedure for creating the set of stored images is also described. The experiment were performed on a total of 120 profiles of 10 persons half of which were used for training. A set of twelve autocorrelation features was used as a feature vector. Three sets of experiments were done. In the first two, 60 randomly chosen training samples were used, while in the third experiment 90 samples were used in the training set. Experiments with varying dimensionality of the training samples are also reported. The best performance (90% accuracy) was achieved when 90 samples were stored in the training set and the dimensionality of training feature vector was four. Comparisons with features derived from moment invariants [?] show that the circular autocorrelations performed better.

Harmon and Hunt [?] present a semi-automatic recognition system for profile posed face recognition by treating the problem as a "waveform" matching problem. The profile photos of 256 males were manually reduced to outline curves by an artist. From these curves, a set of nine fiducial marks such as nose tip, chin, forehead, bridge, nose bottom, throat,

upper lip, mouth and lower lip were automatically identified. The details of how each of these fiducial marks is identified are in [?]. From these fiducial marks, a set of six feature characteristics were derived. These are, protrusion of nose, area right of base line, base angle of profile triangle, wiggle, distances and angles between fiducials. A total of eleven numerical features were extracted from the characteristics mentioned above. After aligning the profiles by using two selected fiducial marks, an Euclidean distance measure is used for measuring the similarity of feature vectors derived from outline profiles. A ranking of most similar faces is obtained by ordering the Euclidean norm. In a subsequent work, Harmon, et al. [?] added images of female subjects and experimented with the same feature vector. By noting that the values of features of a face do not change very much in different images and that faces corresponding to two feature vectors with large Euclidean distance measure between them will be different, a partitioning step is included to improve computational efficiency.

[?] uses the feature extraction methods developped in [?] to create the 11 feature vector components. The 11 feature are reduced to 10 as the Nose Protrusion is highly correlated with 2 other features. The 10-dimensional feature vector is found to provide a high rate of recognition. Classification is done using both classification using Euclidean distances and set partitioning. The set partitioning is used to reduce the number of candidates for inclusion in the Euclidean distance measures and increase performance and diminish computation time. The combination appear to provide a robust system for identifying an unknown with the appropriate file face [?].

[?] is a continuation of the research done in [?] and [?]. The research's aims are to complete a basic understanding of how to achieve automatic identification of human face profiles, to develop robust and economical procedures to use on real-time systems, and to provide the technological framework for further research. The work defines 17 fudicial points which appear to the best combination for face recognition. The method uses minimum Eucledian distance between the unknown and the reference file to determine the correct identification of profile and a thresholding windows for population reduction during the search of reference file. The thresholding windows size is based on the average vector obtain from multiple samples of an individual's profile. In [?], the profiles are obtained from high contrast photography from which transparencies are made, scanned, and digitized. The test set consists of profiles of the same individuals taken at a different setting. The resulting 96% rate of correctness occurs with and without population reduction [?].

Wu and Huang [?] also report a profile based recognition system using an approach similar to that of Harmon and group [?], but significantly different in details. First of all the profile outlines are obtained automatically. B-splines are used to extract six interest points. These are the nose peak, nose bottom, mouth point, chin point, forehead point and the eye

point. A feature vector of dimension 24 is constructed by computing distances between two neighboring points, length, angle between curvature segments joining two adjacent points, etc. Recognition is done by comparing the feature vector extracted from the test image with those stored using the sequential search method using an absolute norm. The stored features are obtained from three instances of a persons profile; in all 18 persons were used for the training phase. The testing data set was generated from the same set of persons used in training, but from a different instance. In the first attempt seventeen of the eighteen test images were correctly recognized. The face image corresponding to the failed case was relearned again (by including the failed image feature vector in the training set). Another instance of this person was correctly recognized using the expanded data set.

Traditional approaches such as Fourier descriptors (FD) have been used for recognizing closed contours. Using p-type FD's [?] that can describe open as well as closed curves, Aibara, et al., [] describe a technique for profile based face recognition. The p-type FD's are derived by discrete Fourier transforming normalized line segments from profiles, are invariant to parallel translation, enlargement/reduction, and are related by simple relation between the original and rotated curves. The training set was generated from three sittings of 90 persons (all males) with the fourth sitting used as the testing data. The p-type FD's (10 coefficients) obtained from three sittings were averaged and used as prototypes. Using 4 coefficients, 65 persons were recognized perfectly. For the full set of 90 test samples, close to 98% accuracy was obtained using 10 coefficients.

5.1 Feature Extraction

5.2 Recognition

6 Forensic Applications

[?] presents an overview of work done using 96 police photographs. The images of the faces are normalized by fixing the pupil distance at 80 pixels. A target face is chosen at random and a neural network is trained to recognize it. The correct face is identified from the 96 100% of the time. With the addition of noise levels of 5% and 10% to the target image, the correct face is found 62.5% and 36.5% of the time. In an experiment using 100 faces, 43 noses are used to train the neural network to find features. The net is able to find the nose feature within 2 pixels using a Euclidean metric. The search area is 10×10. In a profile analysis, a set of 36 profiles are prepared using Fourier descriptors. Cluster analysis is used to group the similarities. The descriptors from the profile can displayed with other data such as height, age, sex, etc. in the form of a histogram or bar-code and may used to increase

search accuracy. Starkey concludes that neural networks will provide a viable solution to identifying criminals photographed at the crime scene. but it is not irrefutable evidence [?].

The initial forensic evidence is often a witness' recall of the culprit's appearance. Verbal descriptions of people's faces very often lack detail. Many times the efforts to assist witnesses' recall using composite techniques are unsuccessful. Face recognition by comparison is very good and witness's are usually called upon to examine mug-shot albums of known criminals. A study was designed to examine the efficiency of an experimental computerized mug-shot retrieval system and album search. Two factors which might be expected to affect retrieval were varied independently. were distinctiveness of target face and the position in the album. Distinctive face are easier to remember than non-distinctive faces. Target faces occurring later in the album are believed to be more difficult to detect than those encountered earlier in the search. The FRAME prototype system with a data base of 1000 faces is used. The faces are rated on a set of 5-point descriptive scales. The scales are derived from the analysis of free descriptions of a different set of faces. The physical measurements corresponding to the features which had been rated were taken from the faces. and tehes were converted to values on 5-point scales using linear regression techniques. The complete record for each face comprises 47 face parameters plus the age. 38 of the parameters are 5-point scale parameters (breadth of face, length of hair, eye color). and 9 are dichotomous parameters code for the presence or absence of facial hair. peculiarities and accessories. Age is coded on 5-point scale. The database consists of 1000 photographs of males between 18-70 taken under controlled conditions. Three photographs were taken. frontal view. profile. and 3/4 view. Four non-distinctive and four distinctive faces were chosen from the set. Eight paid subjects were shown one of the 3/4 view test photographs for 10 sec., provided a detailed description of the photograph. and was assigned randomly to either the computer or album search group. There were four albums in which there were four photographs per page and 250 pages. Each photograph appeared four times within an album. The computer search was performed using the subjects description and ratings and could be changed and repeated up to four times. The computer search retrieval rate for the distinctive faces is 75% and for the non-distinctive faces 69%. The rate for the album searches is 78% for the distinctive faces and 44% for the non-distinctive faces [?].

7 Face Recognition from a Video Sequence

7.1 Segmentation

7.2 Pose Estimation

7.3 The Role of 3-D Models

7.4 Non-rigid Body Motion Analysis

8 Related Issues

In [?] a discussion is presented regarding the recognition of "own-race's" faces being more discriminable than "other-race's faces." Goldstein gives two possible reasons for the discrepancies, psychsocial in which the poor identification results are from the effects of prejudice, unfamiliarity with the class of stimuli, or a variety of other interpersonal reasons, and psychophysical dealing with loss of facial details because of reduced reflectance from dark skin or race-related differences in the variability of facial features. Goldstein includes tables showing the coefficient of variations for different facial features for different races and concludes that poor identification of other races is not a psychophysical problem but more likely a psychsocial one [?]. Using the same data collected in [?], in [?], Goldstein deals with sex differences in face recognition. He finds that in a Japanese population that 65% of the women's facial features are more heterogeneous the men's features. It is found that white women's faces are slightly more variable than the men's but that the overall variation is small. The variation of women's faces is generally slight overall when compared to men's faces [?].

A face must first be detected and located in the scene for identification. The representation and storage format are particular to the recognition system. Each detected faces must be represented in the same format. The representation needs to be compact without the lose of too much information. The type of representation determines the type of matching scheme used in identification. The face may be matched using extracted features or by a wholistic approach from a known data base. The feature set maybe extracted from a frontal view or from a profile. Two competing needs must be balanced in the storage of the face representation: memory usage and the information content. Representations may be in the form of a 2-D image and feature vector. The 2-D form is the simplest but is not compact. The feature vector may be derived from the intensity images or from the profiles. Occasionally a combination of features and intensity data is used to represent faces. Another approach is to represent a face a number or set of numbers such as used by Galton in [?]. In [?] the 2-D image template is used to represent the face. The 2-D template is compacted

as much as possible without losing any information about the face. A set of 64 faces is extracted from images available from different Internet sites. The images, scanned under varying conditions, are not uniform in orientation. There are 10 male and 54 female faces in the set. The minimum resolution needed for human detection is 32×32. The minimum gray scale resolution needed for human detection appears to be 1 bpp if the pattern has a good spatial resolution. Detection of a face can be represented in a few hundred bytes. For identification by a human, the minimum needed for 100% correct identification of the 64 images is $32 \times 32 \times 3$Bpp. Samal concludes that enough information would be contained in $32 \times 32 \times 4$Bpp images for both detection and identification. The space requirement for one face is 512 Bytes and is 512Mbytes for a billion faces [?].

In [?] Nixon uses the Hough transform to achieve facial recognition. The Hough transform is used as it tolerant to noise. The transform locates analytically described shapes by using the magnitude of gradient and the directional information provided by a gradient operator to aid in the recognition process. Two parts of the eye are attractive for recognition of an eye. The almost certain round perimeter of the iris is attractive as detection of a circular shape is an established task in image processing. The perimeter of the eye's sclera is a distinct part and may be employed in detection. The sclera has the advantage in that its shape is reflected by the region below the eyebrows. This allows for eye spacing measurement even when the eyes are closed or the iris is not round due to illness. The analytic shape representing the iris is naturally a circle with the expected gradient directions in each quadrant given the lighter background of the sclera. An ellipse appears the most suitable shape approximating the perimeter of the sclera, but is unsatisfactory for those parts of the eye furthest from the center of the face. The ellipse is tailored for each eye's face center by using an exponential function. Application of the techniques requires generation of a gradient image in which the desired feature is located. [?] uses the Sobel gradient operator. The gradient magnitudes are thresholded using 4 brightness levels to represent the direction of the gradient at that point. The directional information is incorporated into the Hough transform technique. The procedure for locating each eye is constrained to that half of the image. The Hough transform is applied to detect the instance of each shape in a 6 subject data set. The detection of the position of the iris' center from the estimated value has a mean difference of 0.33 pixels. The application of the Hough transform to detect the perimeter of the shape of the region below the eyebrows appears on average to have a spacing 20% larger than the irises' spacing. Using the Hough transform to find the sclera shows the spacing differed on average -1.33 pixels. The results show that it is possible to derive a measurement of the spacing by detection of the position of both the iris, and the shape describing both the perimeter of the sclera and the eyebrows. The measurement by detection of the position of the iris is most accurate.

Detection of the perimeter of the sclera is the most sensitive of the methods. Detection of the position of the eyebrows' provides a measurement of eye spacing which is greater than the other techniques but which may be used when the others can not be discriminated [?].

In [?] the relationship between image quality and recognition of a human face are explored. The task required of observers is to identify one face from a gallery of 35 faces. Using the Modulation Transfer Function Area (MTFA) as a metric to to predict an observers performance in a task requiring the extraction of detailed information from both static and dynamic displays. Performance for an observer is measured by two dependent variables - proportion of correct responses and response time. It was found that as the MTFA_{SQ} becomes moderately large that the facial recognition performance reaches a ceiling which cannot be exceeded. The MTFA metric indicates the extent to which a system response exceeds the minimum observer contrast requirements, average across all spatial frequencies of interest [?].

9 Evaluation of a Face Recognition System

10 Summary and Conclusion