Robust Face Recognition for Data Mining

Brian C. Lovell and Shaokang Chen

Intelligent Real-Time Imaging and Sensing Group

EMI, School of ITEE, The University of Queensland

Australia 4072

{lovell, shaokang}@itee.uq.edu.au

INTRODUCTION

While the technology for mining text documents in large databases could be said to be relatively mature, the same cannot be said for mining other important data types such as speech, music, images and video. Yet these forms of multimedia data are becoming increasingly prevalent on the internet and intranets as bandwidth rapidly increases due to continuing advances in computing hardware and consumer demand. An emerging major problem is the lack of accurate and efficient tools to query these multimedia data directly, so we are usually forced to rely on available metadata such as manual labeling. Currently the most effective way to label data to allow for searching of multimedia archives is for humans to physically review the material. This is already uneconomic or, in an increasing number of application areas, quite impossible because these data are being collected much faster than any group of humans could meaningfully label them — and the pace is accelerating, forming a veritable explosion of non-text data. Some driver applications are emerging from heightened security demands in the 21st century, postproduction of

digital interactive television, and the recent deployment of a planetary sensor network overlaid on the internet backbone.

BACKGROUND

Although they say a picture is worth a thousand words, computer scientists know that the ratio of information contained in images compared to text documents is often much greater than this. Providing text labels for image data is problematic because appropriate labeling is very dependent on the typical queries users will wish to perform, and the queries are difficult to anticipate at the time of labeling. For example, a simple image of a red ball would be best labeled as sports equipment, a toy, a red object, a round object, or even a sphere, depending on the nature of the query. Difficulties with text metadata have led to researchers concentrating on techniques from the fields of Pattern Recognition and Computer Vision that work on the image content itself.

A motivating application and development testbed is the emerging experimental planetary scale sensor web, IrisNet (Gibbons, Karp, Ke, Nath, and Sehan, 2003). IrisNet uses internet connected desktop PCs and inexpensive, off-the-shelf sensors such as Webcams, microphones, temperature, and motion sensors deployed globally to provide a wide-area sensor network. IrisNet is deployed as a service on PlanetLab (www.planet-lab.org), a worldwide collaborative network environment for prototyping next generation internet services initiated by Intel Research and Princeton University that has 177 nodes as of August, 2004. Gibbons, Karp, Ke, Nath, and Sehan envisage a worldwide sensor web in which many users can query, as a single unit, vast quantities of data from thousands or even millions of planetary sensors. IrisNet stores its sensor derived data in a distributed XML schema which is well-suited to describing such hierarchical data as it employs self-

describing tags. Indeed the robust distributed nature of the database can be most readily compared to the structure of the internet DNS naming service.

The authors give an example of IrisNet usage where an ecologist wishes to assess the environmental damage after an oil spill by locating beaches where oil has affected the habitat. The query would be directed toward a coastal monitoring service which collects images from video cameras directed at the coastline. The ecologist would then receive images of the contaminated sites as well at their geographic coordinates. Yet the same coastal monitoring service could be used simultaneously to locate the best beaches for surfing. Moreover, via stored trigger queries, the sensor network could automatically notify the appropriate lifeguard in the event of detecting dangerous rips or the presence of sharks.

A valuable prototype application that could be deployed on IrisNet is wide area person recognition and location services. Such services have existed since the emergence of human society to locate specific persons when they are not in immediate view. For example, in a crowded shopping mall, a mother may ask her child, "Have you seen your sister?" If there were a positive response, this may then be followed by a request to know the time and place of the last sighting, or perhaps by a request to go look for her. Here the mother is using the eyes, face recognition ability, memory persistence, and mobility of the child to perform the search. If the search fails, the mother may then ask the mall manager to give a "lost child" announcement over the public address system. Eventually the police may be asked to employ these human search services on a much wider scale by showing a photograph of the missing child on the television to ask the wider community for assistance in the search.

On the IrisNet the mother could simply upload a photograph of her child from the image store in her mobile phone and the system would efficiently look for the

child in an ever-widening geographic search space until contact was made. Clearly in the case of IrisNet, there is no possibility of humans being employed to identify all the faces captured by the planetary sensor web to support the search, so the task must be automated. Such a service raises inevitable privacy concerns which must be addressed, but the service also has the potential for great public good as in this example of reuniting a worried mother with her lost child.

In addition to person recognition and location services on a planetary sensor web, another interesting commercial application of face recognition is a system to semi-automatically annotate video streams to provide content for digital interactive television. A similar idea was behind the MIT MediaLab Hypersoap project (The Hypersoap Project, Agamanolis and Bove, 1997). In this system, users touch images of objects and people on a television screen to bring up information and advertising material related to the object. For example, a user might select a famous actor and then a page would appear describing the actor, films in which they have appeared, and the viewer might be offered the opportunity to purchase copies of their other films. Automatic face recognition and tracking would greatly simplify the task of labeling the video in post-production — the major cost component of producing such interactive video.

Now we will focus on the crucial technology underpinning such data mining services — automatically recognizing faces in image and video databases.

MAIN THRUST OF THE CHAPTER

Robust Face Recognition

Robust face recognition is a challenging goal because of the gross similarity of all human faces compared to large differences between face images of the same person due to variations in lighting conditions, view point, pose, age, health, and facial expression. An ideal face recognition system should recognize new images of a known face and be insensitive to nuisance variations in image acquisition. Yet, differences between images of the same face (intraclass variation) due to these nuisance variations in image capture are often greater than those between different faces (interclass variation) (Adinj, Moses, and Ulman, 1997), making the task extremely challenging. Most systems work well only with images taken under constrained or laboratory conditions where lighting, pose, and camera parameters are strictly controlled. This requirement is much too strict to be useful in many data mining situations when only few sample images are available such as in recognizing people from surveillance videos from a planetary sensor web or searching historic film archives.

Recent research has been focused on diminishing the impact of nuisance factors on face recognition. Two main approaches have been proposed for illumination invariant recognition. The first is to represent images with features that are less sensitive to illumination change (Yilmaz and Gokmen, 2000, Gao and Leung, 2002) such as using the edge maps of an image. These methods suffer from robustness problems because shifts in edge locations resulting from small rotation or location errors significantly degrade recognition performance. Yilmaz and Gokmen (2000)

proposed using "hills" for face representation; others use derivatives of the intensity (Edelman, Reisfeld, and Yeshurun, 1994, Belhumeur and Kriegman, 1998). No matter what kind of representation is used, these methods assume that features do not change dramatically with variable lighting conditions. Yet this is patently false as edge features generated from shadows may have a significant impact on recognition.

The second main approach is to construct a low dimensional linear subspace for the images of faces taken under different lighting conditions. This method is based on the assumption that images of a convex Lambertian object under variable illumination form a convex cone in the space of all possible images (Belhumeur and Kriegman, 1998). Once again, it is hard for these systems to deal with cast shadows. Furthermore, such systems need several images of the same face taken under different lighting source directions to construct a model of a given face — in data mining applications it is often impossible to obtain the required number of images. Experiments performed by Adinj, Moses, and Ulman (1997) show that even with the best image representations using illumination insensitive features and the best distance measurement, the misclassification rate is often more than 20%.

Table 1: Problems with Existing Face Recognition Technology

- Overall accuracy, particularly on large databases
- Sensitivity to changes in lighting, camera angle, pose
- Computational load of searches

As for expression invariant face recognition, this is still an open problem for machine recognition and is also quite a difficult task for humans. The approach

adopted in Beymer and Poggio (1996) and Black, Fleet, and Yacoob (2000) is to morph images to be the same expression as the one used for training. A problem is that not all images can be morphed correctly. For example an image with closed eyes cannot be morphed to a standard image because of the lack of texture inside the eyes. Liu, Chen, and Kumar (2001) proposed using optical flow for face recognition with facial expression variations. However, it is hard to learn the motions within the feature space to determine the expression changes, since the way one person express a certain emotion is normally somewhat different from others. These methods also suffer from the need to have large numbers of example images for training.

Table 2: Data Mining Applications for Face Recognition

- Person recognition and location services on a planetary wide sensor net
- Recognizing faces in a crowd from video surveillance
- Searching for video or images of selected persons in multimedia databases
- Forensic examination of multiple video streams to detect movements of certain persons
- Automatic annotation and labeling of video streams to provide added value for digital interactive television

Mathematical Basis for Face Recognition Technologies

Most face recognition systems are based on one of the following methods:

1. Direct Measurement of Facial Features
2. Principal Components Analysis or "Eigenfaces" (Turk and Pentland, 1991)
3. Fisher Linear Discriminant Function (Liu and Wechsler, 1998)

Early forms of face recognition were based on Method 1 with direct measurement of features such as width of the nose, spacing between the eyes, etc. These measurements were frequently performed by hand using calipers. Many modern systems are based on either of Methods 2 or 3 which are better suited to computer automation. Here we briefly describe the principles behind one of the most popular methods — Principal Components Analysis (PCA), also known as "eigenfaces," as originally popularized by Turk and Pentland (1991). The development assumes a basic background in linear algebra.

Principal Components Analysis

PCA is a second-order method for finding a linear representation of faces using only the covariance of the data. It determines the set of orthogonal components (feature vectors) which minimizes the reconstruction error for a given number of feature vectors. Consider the face image set $I = [I_1, I_2, \cdots, I_n]$, where I_i is a $p \times q$ pixel image, $i \in [1 \cdots n]$, $p, q, n \in Z^+$, the average face of the image set is defined by the matrix:

$$\Psi = \frac{1}{n} \sum_{k=1}^{n} I_i \,. \qquad (1)$$

Note that face recognition is normally performed on grayscale (i.e., black and white) face images rather than color. Colors, and skin color tones in particular, are frequently used to aid face detection and location within the image stream (Rein-Lien, Abdel-Mottaleb, and Jain, 2002). We assume additionally that the face images are pre-processed by scaling, rotation, eye centre alignment, and background suppression so that averaging is meaningful. Now normalizing each image by subtracting the average face, we have the normalized difference image matrix:

$$\tilde{D}_i = I_i - \Psi . \tag{2}$$

Unpacking $\tilde{D}_i$ row-wise, we form the N $(N = p \times q)$ dimensional column vector d_i. We define the covariance matrix C of the normalized image set $D = [d_1, d_2, \cdots d_n]$ corresponding to the original face image set I by:

$$C = \sum_{i=1}^{n} d_i d_i^T = DD^T . \tag{3}$$

An eigendecomposition of C yields eigenvalues λ_i and eigenvectors u_i which satisfy:

$$Cu_i = \lambda_i u_i , \tag{4}$$

$$C = DD^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T , \tag{5}$$

where $i \in [1 \cdots N]$.

In practice, N is so huge that eigenvector decomposition is computationally impossible. Indeed for even a small image of 100×100 pixels, C is a $10{,}000 \times 10{,}000$ matrix. Fortunately, the following shortcut lets us bypass direct decomposition of C.

We consider decompositions of $C' = D^T D$ instead of $C = DD^T$. Singular value decomposition of D gives us

$$D = USV^T \tag{6}$$

where $U^{[N\times N]}$ and $V^{[n\times n]}$ are unitary and $S^{[Nxn]}$ is diagonal. Without loss of generality, assume the diagonal elements of $S = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$ are sorted such that $\sigma_1 > \sigma_2 > \cdots > \sigma_n$ where the σ_i are known as the singular values of D. Then

$$C = DD^T = USV^T VS^T U^T = US^2 U^T = \sum_{i=1}^{n} \sigma_i^2 u_i u_i^T \tag{7}$$

where $S^{2[N\times N]} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2, 0, \cdots 0)$. Thus, only the first n singular values are non zero.

Comparing (7) with (5), we see that the squares of the singular values give us the eigenvalues of C (i.e., $\lambda_i = \sigma_i^2$) and the columns of U are the eigenvectors. Now consider a similar derivation for $C^{'}$.

$$C' = D^T D = VS^T U^T USV^T = VS^2 V^T = \sum_{i=1}^{n} \sigma_i^2 v_i v_i^T \tag{8}$$

where $S^{2[n\times n]} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$. Comparing (7) and (8) we see that the singular values are identical, so the squares of the singular values yield the eigenvalues of C. The eigenvectors of C can be obtained from the eigenvectors of C', which are the columns of V, by rearranging (6) as follows:

$$U = DVS^{-1} \tag{9}$$

which can be expressed alternatively by

$$u_i = \frac{1}{\sqrt{\lambda_i}} Dv_i, \tag{9}$$

where $i = [1 \cdots n]$. Thus by performing an eigenvector decomposition on the small matrix $C'^{[n\times n]}$, we efficiently obtain both the eigenvalues and eigenvectors of the very large matrix $C^{[N\times N]}$. In the case of a database of 100×100 pixel face images of size 30, by using this shortcut, we need only decompose a 30×30 matrix instead of a $10{,}000\times 10{,}000$ matrix!

The eigenvectors of C are often called the eigenfaces and are shown as images in Figure 1. Being the columns of a unitary matrix, the eigenfaces are orthogonal and efficiently describe (span) the space of variation in faces. Generally, we select a small subset of $m < n$ eigenfaces to define a reduced dimensionality facespace that yields highest recognition performance on unseen examples of faces. For good recognition performance the required number of eigenfaces, m, is typically chosen to be of the order of 6 to 10.

Thus in PCA recognition each face can be represented by just a few components by subtracting out the average face and then calculating principal components by projecting the remaining difference image onto the m eigenfaces. Simple methods such as nearest neighbors are normally used to determine which face best matches a given face.

Figure 1. Typical set of eigenfaces as used for face recognition. Leftmost image is average face.

Robust PCA Recognition

The authors have developed Adaptive Principal Component Analysis (APCA) to improve the robustness of PCA to nuisance factors such as lighting and expression (Chen and Lovell, 2003 and 2004). In the APCA method, we first apply PCA. Then we rotate and warp the facespace by whitening and filtering the

eigenfaces according to overall covariance, between-class, and within-class covariance to find an improved set of eigenfeatures. Figure 2 shows the large improvement in robustness to lighting angle. The proposed APCA method allows us to recognize faces with high confidence even if they are half in shadow. Figure 3 shows significant recognition performance gains over standard PCA when both changes in lighting and expression are present.

Critical Issues of Face Recognition Technology

Despite the huge number of potential applications for reliable face recognition, the need for such search capabilities in multimedia data mining, and the great strides made in recent decades, there is still much work to do before these applications become routine

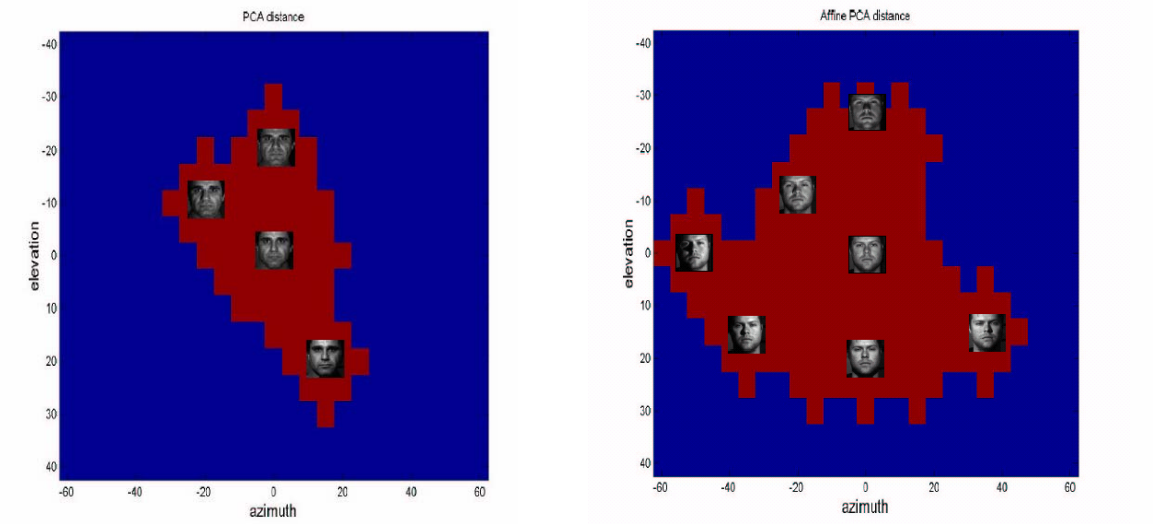

Figure 2. Contours of 95% recognition performance for the original PCA and the proposed APCA method against lighting elevation and azimuth.

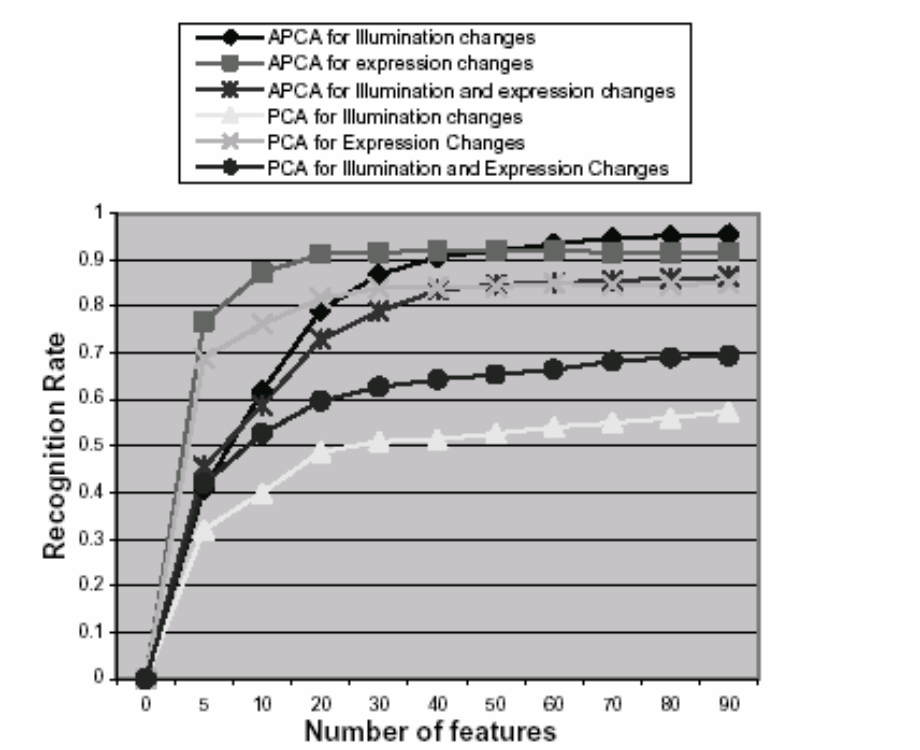

Figure 3. Recognition rates for APCA and PCA versus number of eigenfaces with variations in lighting and expression from Chen and Lovell (2003).

Table 3: A Summary of Critical Issues of Face Recognition Technologies

Privacy Concerns

It is clear that personal privacy may be reduced with the widespread adoption of face recognition technology. However, since 911, concerns about privacy have taken a back seat to concerns about personal security. Governments are under intense pressure to introduce stronger security measures. Unfortunately government's current need for biometric technology does nothing to improve performance in the short term and may actually damage uptake in the medium term due to unrealistic expectations.

Computational Efficiency

Face recognition can be computationally very intensive for large databases. This is a serious impediment for multimedia datamining.

Accuracy on Large Databases

Studies indicate that recognition error rates of the order of 10% are the best that can be obtained on large databases. This error rate sounds rather high, but trained humans do no better and are much slower at searching.

Sensitivity to Illumination and Other Changes

Changes in lighting, camera angle, and facial expression can greatly affect recognition performance.

Inability to Cope with Multiple Head Poses

Very few systems can cope with non-frontal views of the face. Some researchers propose 3D recognition systems using stereo cameras for real-time applications, but these are not suitable for data mining.

Ability to Scale

While a laboratory system may work quite well on 20 or 30 faces, it is not clear that these systems will scale to huge face databases as required for many security applications such as detecting faces of known criminals in a crowd or the person locator service on the planetary sensor web.

FUTURE TRENDS

Face recognition and other biometric technologies are coming of age due to the need to address heightened security concerns in the 21st century. Privacy concerns that have hindered public acceptance of these technologies in the past are now yielding to society's need for increased security while maintaining a free society. Apart from the demands from the security sector, there are many applications for the technology in other areas of data mining. The performance and robustness of systems will increase significantly as more researcher effort is brought to bear. In recent real-time systems there is much interest in 3D reconstruction of the head from multiple camera angles, but in data mining the focus must remain on reliable recognition from single photos.

CONCLUSION

It has been argued that by the end of the 20th century computers were very capable of handling text and numbers and that in the 21st century computers will have to be able to cope with raw data such as images and speech with much the same facility. The explosion of multimedia data on the internet and the conversion of all information to digital formats (music, speech, television) is driving the demand for advanced multimedia search capabilities, but the pattern recognition technology is mostly unreliable and slow. Yet, the emergence of handheld computers with built-in speech and handwriting recognition ability, however primitive, is a sign of the changing times. The challenge for researchers is to produce pattern recognition algorithms, such as face recognition, reliable and fast enough for deployment on data spaces of a planetary scale.

REFERENCES

Adinj, Y., Moses, Y. and Ullman, S. (1997) "Face recognition: The problem of compensation for changes in illumination direction", IEEE PAMI, 19(4), 721-732.

Agamanolis, Stefan and Bove, V. Michael Jr. (1997), ``Multi-Level Scripting for Responsive Multimedia,'' IEEE Multimedia, 4(4), 40-50.

Belhumeur, P., and Kriegman, D. (1998), "What Is the Set of Images of an Object under All Possible Illumination Conditions", Int'l J.Computer Vision, 28(3), 245-260.

Beymer, D., and Poggio, T. (1995), "Face Recognition from One Example View", Proc. Int'l Conf. of Comp. Vision, 500-507.

Black, M. J., Fleet, D. J. and Yacoob, Y. (2000), "Robustly estimating Changes in Image Appearance", Computer Vision and Image Understanding, 78(1), 8-31.

Chen, Shaokang and Lovell, Brian C. (2004), “Illumination and Expression Invariant Face Recognition with One Sample Image,” Proceedings of the International Conference on Pattern Recognition, Cambridge, August 23-26.

Chen, Shaokang and Lovell, Brian C. (2003), "Face Recognition with One Sample Image per Class," Proceedings of ANZIIS2003, Sydney, December 10-12, 83-88.

Chen, Shaokang, Lovell, Brian C., and Sun, S. (2002), “Face recognition with APCA in variant illuminations,” Proc of WOSPA2002, December 17-18, Brisbane, 9-12.

Edelman, S., Reisfeld, D. and Yeshurun, Y.(1994), "A System for Face Recognition that Learns from Examples", Proc. European Conf. Computer Vision Springer-Verlag, 787-791.

Feraud, R., Bernier, O. Viallet, J.E. and Collobert, M. (2000), "A fast and accurate face detector for indexation of face images," Proc. Fourth IEEE International Conference on Automatic Face and Gesture Recognition, March 28-30,77 - 82.

Gao, Yongsheng and Leung, Maylor K.H.(2002), "Face Recognition Using Line Edge Map", IEEE PAMI. 24(6), June, 764-779.

Georghiades, A.S., Belhumeur, P.N. and Kriegman, D.J. (2001), "From few to many: illumination cone models for face recognition under variable lighting and pose," IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6), June, 643-660.

Gibbons, P.B., Karp, B., Ke, Y., Nath, S., and Sehan S (2003), "IrisNet: An Architecture for a Worldwide Sensor Web," Pervasive Computing, 2(4), Oct – Dec, 22-33.

Jian Yang, Zhang, D., Frangi, A. F. and Jing-Yu, Yang (2004), "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(1), Jan, pp 131-137.

Liu, X. M., Chen, T. and Kumar, B.V.K.V. (2003), "Face Authentication for Multiple Subjects Using Eigenflow", Pattern Recognition, Special issue on Biometric, 36(2), 313-328.

Li, Y, Goshtasby, A., and Garcia, O.(2000), "Detecting and tracking human faces in videos," Proc. 15th Int'l Conference on Pattern Recognition, 2000, 3-7 Sept, 1, 807-810.

Liu, Chengjun and Wechsler, Harry (1998), "Evolution of Optimal Projection Axes (OPA) for Face Recognition", Third IEEE International Conference on Automatic face and Gesture Recognition, FG'98, Nara, Japan, April 14-16, 282-287.

Ming-Hsuan, Yang, Kriegman, D.J., and Ahuja, N. (2002) "Detecting faces in images: a survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Jan, 24(1), 34-58.

Rein-Lien, Hsu, Abdel-Mottaleb, M. and Jain, A. K. (2002) "Face detection in color images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), May, 696-706.

Swets, Daniel L. and Weng, John (1996), "Using discriminant eigenfeatures for image retrieval", IEEE Transactions on Pattern Analysis and Machine Intelligence, 18(8), 831-836.

The Hypersoap Project, http://www.media.mit.edu/hypersoap/, [online], [last visited 6-Feb-2004].

Turk M. A., and Pentland, A. P. (1991), "Eigenfaces for recognition", Journal of Cognitive Neuroscience, 3(1), 71-86.

Zhao, L., and Yang, Y. H. (1999), “Theoretical Analysis of Illumination in PCA-based Vision Systems”, Pattern Recognition, 32, 547-564.

Yilmaz, A. and Gokmen, M. (2000), "Eigenhill vs. eigenface and eigenedge", In Procs of International Conference Pattern Recognition, Barcelona, Spain, 827-830.

TERMS AND THEIR DEFINITION

Biometric: A measurable, physical characteristic or personal behavioral trait used to recognize the identity, or verify the claimed identity, of an enrollee. A biometric identification system identifies a human from a measurement of a physical feature or repeatable action of the individual (for example, hand geometry, retinal scan, iris scan, fingerprint patterns,

facial characteristics, DNA sequence characteristics, voice prints, and hand written signature).

Computer Vision: Using computers to analyze images and video streams and extract meaningful information from them in a similar way to the human vision system. It is related to artificial intelligence and image processing and is concerned with computer processing of images from the real world to recognize features present in the image.

Eigenfaces: Another name for face recognition via principal components analysis.

Face Space: The vector space spanned by the eigenfaces.

Head Pose: Position of the head in 3D space including head tilt, and rotation.

Metadata: Labeling, information describing other information.

Pattern Recognition: Pattern Recognition is the ability to take in raw data, such as images, and take action based on the category of the data.

Principal Components Analysis: Principal components analysis (PCA) is a method that can be used to simplify a dataset. It is a transform that chooses a new coordinate system for the data set, such, that that greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis and so on. PCA can be used for reducing dimensionality. PCA is also called the Karhunen-Loève transform or the Hotelling transform.

Robust: The opposite of Brittle; this can be said of a system that has the ability to recover gracefully from the whole range of exceptional inputs and situations in a given environment. Also has the connotation of elegance in addition to careful attention to detail.