Personalizing Smart Environments: Face Recognition for Human Interaction

Alex Pentland and Tanzeem Choudhury
The Media Laboratory
Massachusetts Institute of Technology,
20 Ames St., Cambridge, MA 02139
{sandy,tanzeem}@media.mit.edu, http://www.media.mit.edu/~pentland

October 8, 1999

Abstract

A smart environment is one that is able to identify people, interpret their actions, and react appropriately. Thus, one of the most important building blocks of smart environments is a person identification system. Face recognition devices are ideal for such systems, since they have recently become fast, cheap, unobtrusive, and, when combined with voice-recognition, are very robust against changes in the environment. Moreover, since humans primarily recognize each other by their faces and voices, they feel comfortable interacting with an environment that does the same. We present a brief summary of the history and mathematical framework of face recognition, the current state of the art, and present-day commercial systems. We then describe developments towards future applications: building interactive smart environments, augmenting human senses, skills and memory with wearable recognition technology, and ultimately making computers so usable, portable and intuitive that they become ubiquitous — the so-called "fourth generation" of computing.

1 Introduction

Smart environments, wearable computers, and ubiquitous computing in general are thought to be the coming 'fourth generation' of computing and information technology [1, 2, 3]. Because these devices will be everywhere — clothes, home, car, and office, their economic impact and cultural significance are expected to dwarf previous generations of computing. At a minimum, they are among the most exciting and economically important research areas in information technology and computer science.

However, before this new generation of computing can be widely deployed we must invent new methods of interaction that don't require a keyboard or mouse — there will be too many small computers to instruct them all individually. To win wide consumer acceptance such interactions must be friendly and personalized (no one likes being treated like just another cog in a machine!), which implies that next-generation interfaces will be aware of the people in their immediate environment and at a minimum know who they are.

1.1 Why Face Recognition?

Given the requirement for determining people's identity, the obvious question is what technology is best suited to supply this information? There are many different identification technologies available, many of which have been in wide-spread commercial use for years. The most common person verification and identification methods today are Password/PIN (Personal Identification Number) systems, and Token

systems (such as your driver's license). Because such systems have trouble with forgery, theft, and lapses in users' memory, there has developed considerable interest in biometric identification systems, which use pattern recognition techniques to identify people using their physiological characteristics. Fingerprints are a classic example of a biometric; newer technologies include retina and iris recognition.

While appropriate for bank transactions and entry into secure areas, such technologies have the disadvantage that they are intrusive both physically and socially. They require the user to position their body relative to the sensor, and then pause for a second to 'declare' themselves. This 'pause and declare' interaction is unlikely to change because of the fine-grain spatial sensing required. Moreover, there is a 'oracle-like' aspect to the interaction: since people can't recognize other people using this sort of data, these types of identification do not have a place in normal human interactions and social structures.

While the 'pause and present' interaction and the oracle-like perception are useful in high-security applications (they make the systems look more accurate), they are exactly the opposite of what is required when building a store that recognizes its best customers, or an information kiosk that remembers you, or a house that knows the people who live there. Face recognition from video and voice recognition have a natural place in these next-generation smart environments — they are unobtrusive (able to recognize at a distance without requiring a 'pause and present' interaction), are usually passive (do not require generating special electro-magnetic illumination), do not restrict user movement, and are now both low-power and inexpensive. Perhaps most important, however, is that humans identify other people by their face and voice, therefore are likely to be comfortable with systems that use face and voice recognition.

2 History and Mathematical Framework

Twenty years ago the problem of face recognition was considered among the hardest in Artificial Intelligence (AI) and computer vision. Surprisingly, however, over the last decade there have been a series of successes that have made the general person identification enterprise appear not only technically feasible but also economically practical.

The apparent tractability of face recognition problem combined with the dream of smart environments has produced a huge surge of interest from both funding agencies and from researchers themselves. It has also spawned several thriving commercial enterprises. There are now several companies that sell commercial face recognition software that is capable of high-accuracy recognition with databases of over 1,000 people.

These early successes came from the combination of well-established pattern recognition techniques with a fairly sophisticated understanding of the image generation process. In addition, researchers realized that they could capitalize on regularities that are peculiar to people, for instance, that human skin colors lie on a one-dimensional manifold (with color variation primarily due to melanin concentration), and that human facial geometry is limited and essentially 2-D when people are looking toward the camera. Today, researchers are working on relaxing some of the constraints of existing face recognition algorithms to achieve robustness under changes in lighting, aging, rotation-in-depth, expression and appearance (beard, glasses, makeup) — problems that have partial solution at the moment.

2.1 The Typical Representational Framework

The dominant representational approach that has evolved is descriptive rather than generative. Training images are used to characterize the range of 2-D appearances of objects to be recognized. Although initially very simple modeling methods were used, the dominant method of characterizing appearance has fairly quickly become estimation of the probability density function (PDF) of the image data for the target class.

For instance, given several examples of a target class Ω in a low-dimensional representation of the image data, it is straightforward to model the probability distribution function $P(\mathbf{x}|\Omega)$ of its image-level features

$\mathbf{x}$ as a simple parametric function (e.g., a mixture of Gaussians), thus obtaining a low-dimensional, computationally efficient *appearance model* for the target class.

Once the PDF of the target class has been learned, we can use Bayes' rule to perform maximum *a posteriori* (MAP) detection and recognition. The result is typically a very simple, neural-net-like representation of the target class's appearance, which can be used to detect occurrences of the class, to compactly describe its appearance, and to efficiently compare different examples from the same class. Indeed, this representational framework is so efficient that some of the current face recognition methods can process video data at 30 frames per second, and several can compare an incoming face to a database of thousands of people in under one second — and all on a standard PC!

2.2 Dealing with the Curse of Dimensionality

To obtain an 'appearance-based' representation, one must first transform the image into a low-dimensional coordinate system that preserves the general perceptual quality of the target object's image. This transformation is necessary in order to address the 'curse of dimensionality'. The raw image data has so many degrees of freedom that it would require millions of examples to learn the range of appearances directly.

Typical methods of dimensionality reduction include Karhunen-Loève transform (KLT) (also called Principal Components Analysis (PCA)) or the Ritz approximation (also called 'example-based representation'). Other dimensionality reduction methods are sometimes also employed, including sparse filter representations (e.g., Gabor Jets, Wavelet transforms), feature histograms, independent components analysis, and so forth.

These methods have in common the property that they allow efficient characterization of a low-dimensional subspace with the overall space of raw image measurements. Once a low-dimensional representation of the target class (face, eye, hand, etc.) has been obtained, standard statistical parameter estimation methods can be used to learn the range of appearance that the target exhibits in the new, low-dimensional coordinate system. Because of the lower dimensionality, relatively few examples are required to obtain a useful estimate of either the PDF or the inter-class discriminant function.

An important variation on this methodology is *discriminative models*, which attempt to model the differences between classes rather than the classes themselves. Such models can often be learned more efficiently and accurately than when directly modeling the PDF. A simple linear example of such a difference feature is the Fisher discriminant. One can also employ discriminant classifiers such as Support Vector Machines (SVM) which attempt to maximize the margin between classes.

3 Person Identification via Face Recognition

The current literature on face recognition contains thousands of references, most dating from the last few years. For an exhaustive survey of face analysis techniques the reader is referred to Chellappa et al. [4], and for current research the reader is referred to the IEEE Conferences on Automatic Face and Gesture Recognition.

Research on face recognition goes back to the earliest days of AI and computer vision. Rather than attempting to produce an exhaustive historical account, our focus will be on the early efforts that had the greatest impact on the community (as measured by, e.g., citations), and those few current systems that are in wide-spread use or have received extensive testing.

3.1 History of Face Recognition

The subject of face recognition is as old as computer vision, both because of the practical importance of the topic and theoretical interest from cognitive scientists. Despite the fact that other methods of identification (such as fingerprints, or iris scans) can be more accurate, face recognition has always

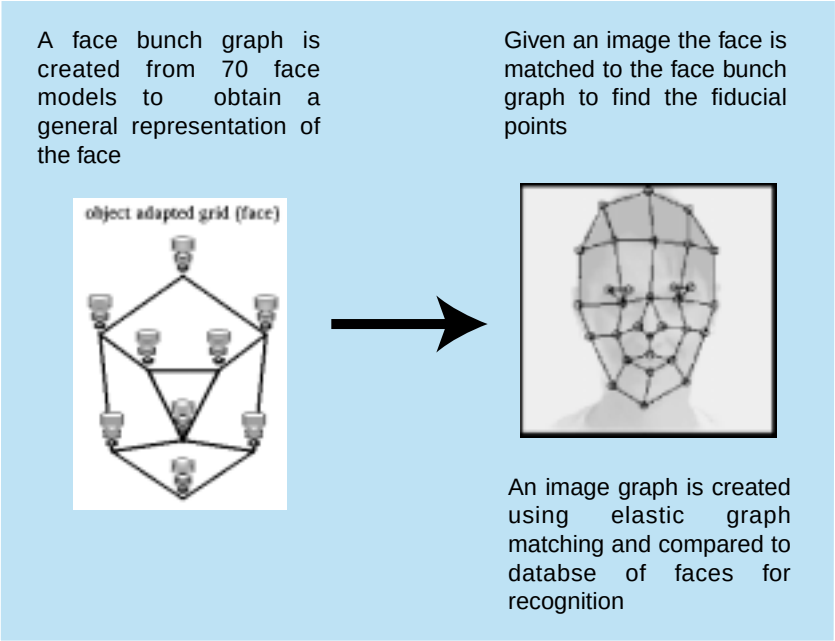

Figure 1: Face Recognition using Elastic Graph Matching

remains a major focus of research because of its non-invasive nature and because it is people's primary method of person identification.

Perhaps the most famous early example of a face recognition system is due to Kohonen [5], who demonstrated that a simple neural net could perform face recognition for aligned and normalized face images. The type of network he employed computed a face description by approximating the eigenvectors of the face image's autocorrelation matrix; these eigenvectors are now known as 'eigenfaces.'

Kohonen's system was not a practical success, however, because of the need for precise alignment and normalization. In following years many researchers tried face recognition schemes based on edges, inter-feature distances, and other neural net approaches. While several were successful on small databases of aligned images, none successfully addressed the more realistic problem of large databases where the location and scale of the face is unknown.

Kirby and Sirovich (1989) [6] later introduced an algebraic manipulation which made it easy to directly calculate the eigenfaces, and showed that fewer than 100 were required to accurately code carefully aligned and normalized face images. Turk and Pentland (1991) [7] then demonstrated that the residual error when coding using the eigenfaces could be used both to detect faces in cluttered natural imagery, and to determine the precise location and scale of faces in an image. They then demonstrated that by coupling this method for detecting and localizing faces with the eigenface recognition method, one could achieve reliable, real-time recognition of faces in a minimally constrained environment. This demonstration that simple, real-time pattern recognition techniques could be combined to create a useful system sparked an explosion of interest in the topic of face recognition.

3.2 Current State of the Art

By 1993 there were several algorithms claiming to have accurate performance in minimally constrained environments. To better understand the potential of these algorithms, DARPA and the Army Research

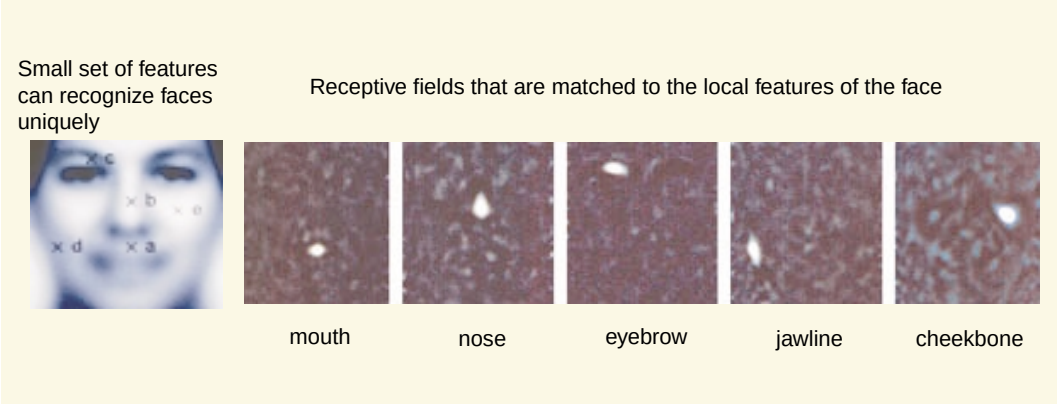

Figure 2: Face Recognition using Local Feature Analysis

Laboratory established the FERET program with the goals of both evaluating their performance and encouraging advances in the technology [8].

At the time of this writing, there are three algorithms that have demonstrated the highest level of recognition accuracy on large databases (1196 people or more) under double-blind testing conditions. These are the algorithms from University of Southern California (USC) [9], University of Maryland (UMD) [10], and the MIT Media Lab [11]. All of these are participants in the FERET program. Only two of these algorithms, from USC and MIT, are capable of both minimally constrained detection and recognition; the others require approximate eye locations to operate. A fourth algorithm that was an early contender, developed at Rockefeller University [12], dropped from testing to form a commercial enterprise. The MIT and USC algorithms have also become the basis for commercial systems.

The MIT, Rockefeller, and UMD algorithms all use a version of the eigenface transform followed by discriminative modeling. The UMD algorithm uses a linear discriminant, while the MIT system, seen in Figure 3, employs a quadratic discriminant. The Rockefeller system, seen in Figure 2, uses a sparse version of the eigenface transform, followed by a discriminative neural network. The USC system, seen in Figure 1, in contrast, uses a very different approach. It begins by computing Gabor 'jets' from the image, and then does a 'flexible template' comparison between image descriptions using a graph-matching algorithm.

The FERET database testing employs faces with variable position, scale, and lighting in a manner consistent with mugshot or driver's license photography. On databases of under 200 people and images taken under similar conditions, all four algorithms produce nearly perfect performance. Interestingly, even simple correlation matching can sometimes achieve similar accuracy for databases of only 200 people [8]. This is strong evidence that any new algorithm should be tested with at databases of at least 200 individuals, and should achieve performance over 95% on mugshot-like images before it can be considered potentially competitive.

In the larger FERET testing (with 1166 or more images), the performance of the four algorithms is similar enough that it is difficult or impossible to make meaningful distinctions between them (especially if adjustments for date of testing, etc., are made). On frontal images taken the same day, typical first-choice recognition performance is 95% accuracy. For images taken with a different camera and lighting, typical performance drops to 80% accuracy. And for images taken one year later, the typical accuracy is approximately 50%. Note that even 50% accuracy is 600 times chance performance.

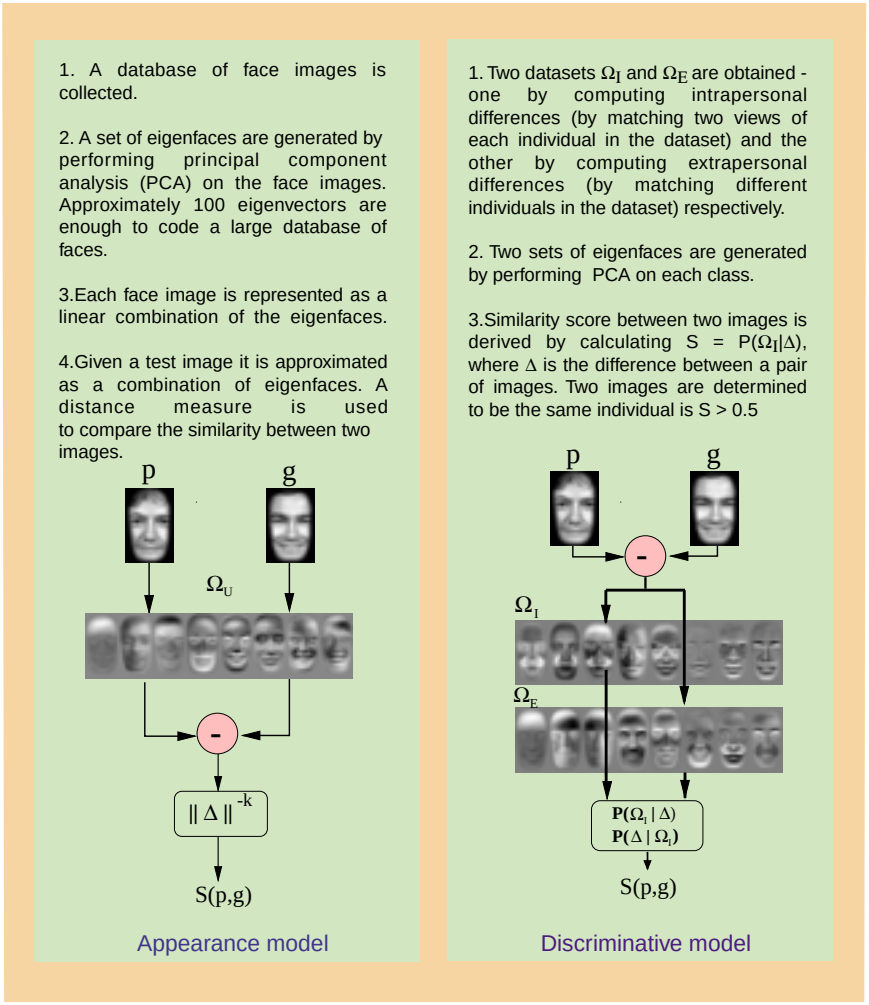

Figure 3: Face Recognition using Eigenfaces

3.3 Commercial Systems and Applications

Currently, several face-recognition products are commercially available. Algorithms developed by the top contenders of the FERET competition are the basis of some of the available systems; others were developed outside of the FERET testing framework. While it is extremely difficult to judge, three systems — Visionics, Viisage, and Miros — seem to be the current market leaders in face recognition.

Visionics' FaceIt face recognition software is based on the Local Feature Analysis algorithm developed at Rockefeller University. FaceIt is now being incorporated into a Close Circuit Television (CCTV) anti-crime system called 'Mandrake' in United Kingdom. This system searches for known criminals in video acquired from 144 CCTV camera locations. When a match occurs a security officer in the control room is notified.

Viisage, another leading face-recognition company, uses the eigenface-based recognition algorithm developed at the MIT Media Laboratory. Their system is used in conjunction with identification cards (e.g., driver's licenses and similar government ID cards) in many US states and several developing nations.

Miros uses neural network technology for their TrueFace face recognition software. TrueFace is used

by Mr. Payroll for their check cashing system, and has been deployed at casinos and similar sites in many US states.

4 Novel Applications of Face Recognition Systems

Face recognition systems are no longer limited to identity verification and surveillance tasks. Growing numbers of applications are starting to use face-recognition as the initial step towards interpreting human actions, intention, and behavior, as a central part of next-generation smart environments. Many of the actions and behaviors humans display can only be interpreted if you also know the person's identity, and the identity of the people around them. Examples are a valued repeat customer entering a store, or behavior monitoring in an eldercare or childcare facility, and command-and-control interfaces in a military or industrial setting. In each of these applications identity information is crucial in order to provide machines with the background knowledge needed to interpret measurements and observations of human actions.

4.1 Face Recognition for Smart Environments

Researchers today are actively building smart environments (i.e. visual, audio, and haptic interfaces to environments such as rooms, cars, and office desks) [1, 2]. In these applications a key goal is usually to give machines perceptual abilities that allow them to function naturally with people — to recognize the people and remember their preferences and peculiarities, to know what they are looking at, and to interpret their words, gestures, and unconscious cues such as vocal prosody and body language. Researchers are using these perceptually-aware devices to explore applications in health care, entertainment, and collaborative work.

Recognition of facial expression is an important example of how face recognition interacts with other smart environment capabilities. It is important that a *smart* system knows whether the user looks impatient because information is being presented too slowly, or confused because it is going too fast — facial expressions provide cues for identifying and distinguishing between these different states. In recent years much effort has been put into the area of recognizing facial expression, a capability that is critical for a variety of human-machine interfaces, with the hope of creating a person-independent expression recognition capability. While there are indeed similarities in expressions across cultures and across people, for anything but the most gross facial expressions analysis must be done relative to the person's normal facial rest state — something that definitely isn't the same across people. Consequently, facial expression research has so far been limited to recognition of a few discrete expressions rather than addressing the entire spectrum of expression along with its subtle variations. Before one can achieve a really useful expression analysis capability one must be able to first recognize the person, and tune the parameters of the system to that specific person.

4.2 Wearable Recognition Systems

When we build computers, cameras, microphones and other sensors into a person's clothes, the computer's view moves from a passive third-person to an active first-person vantage point (see Figure 4) [3]. These wearable devices are able to adapt to a specific user and to be more intimately and actively involved in the user's activities. The field of wearable computing is rapidly expanding, and just recently became a full-fledged Technical Committee within the IEEE Computer Society. Consequently, we can expect to see rapidly-growing interest in the largely-unexplored area of first-person image interpretation.

Face recognition is an integral part of wearable systems like memory aides, remembrance agents, and context-aware systems. Thus there is a need for many future recognition systems to be integrated with the user's clothing and accessories. For instance, if you build a camera into your eyeglasses, then face recognition software can help you remember the name of the person you are looking

Figure 4: Wearable Face Recognition System

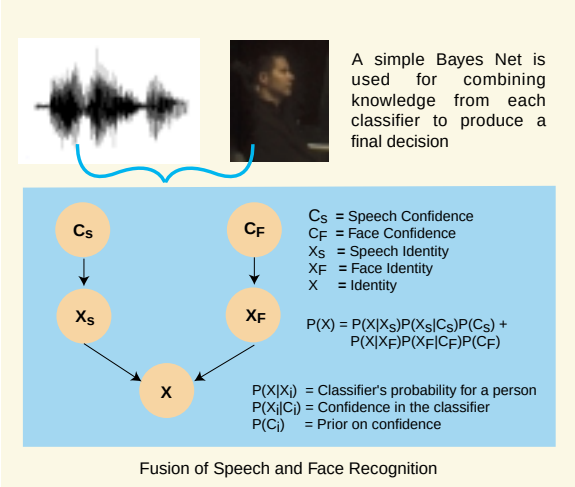

Figure 5: Multi-modal Person Recognition System

at by whispering their name in your ear. Such devices are beginning to be tested by the US Army for use by border guards in Bosnia, and by researchers at the University of Rochester's Center for Future Health for use by Alzheimer's patients (see http://wearables.www.media.mit.edu/projects/wearables & http://www.futurehealth.rochester.edu).

5 Future of Face Recognition Technology

Face recognition systems used today work very well under constrained conditions, although all systems work much better with frontal mug-shot images and constant lighting. All current face recognition algorithms fail under the vastly varying conditions under which humans need to and are able to identify other people. Next generation person recognition systems will need to recognize people in real-time and in much less constrained situations.

We believe that identification systems that are robust in natural environments, in the presence of noise and illumination changes, cannot rely on a single modality, so that fusion with other modalities is

essential (see Figure 5). Technology used in smart environments has to be unobtrusive and allow users to act freely. Wearable systems in particular require their sensing technology to be small, low powered and easily integrable with the user's clothing. Considering all the requirements, identification systems that use face recognition and speaker identification seem to us to have the most potential for wide-spread application.

Cameras and microphones today are very small, light-weight and have been successfully integrated with wearable systems. Audio and video based recognition systems have the critical advantage that they use the modalities humans use for recognition. Finally, researchers are beginning to demonstrate that unobtrusive audio-and-video based person identification systems can achieve high recognition rates without requiring the user to be in highly controlled environments [13].

6 Conclusion

Face recognition technology has come a long way in the last twenty years. Today, machines are able to automatically verify identity information for secure transactions, for surveillance and security tasks, and for access control to buildings etc. These applications usually work in controlled environments and recognition algorithms can take advantage of the environmental constraints to obtain high recognition accuracy. However, next generation face recognition systems are going to have widespread application in smart environments — where computers and machines are more like helpful assistants.

To achieve this goal computers must be able to reliably identify nearby people in a manner that fits naturally within the pattern of normal human interactions. They must not require special interactions and must conform to human intuitions about when recognition is likely. This implies that future smart environments should use the same modalities as humans, and have approximately the same limitations. These goals now appear in reach — however, substantial research remains to be done in making person recognition technology work reliably, in widely varying conditions using information from single or multiple modalities.

References

[1] M. Weiser, "The computer for the 21st century," *Scientific American*, vol. 265, no. 3, pp. 66–76, 1991.

[2] A. Pentland, "Smart rooms, smart clothes," *Scientific American*, vol. 274, no. 4, pp. 68–76, 1996.

[3] A. Pentland, "Wearable intelligence," *Scientific American Presents*, vol. 9, no. 4, pp. 90–95, 1998.

[4] R. Chellappa, C. Wilson, and S. Sirohev, "Human and machine recognition of faces: A survey," in *Proceedings of IEEE*, May 1995, vol. 83, pp. 705–740.

[5] T. Kohonen, *Self-organization and Associative Memory*, Springer-Verlag, Berlin, 1989.

[6] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[7] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[8] P. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[9] L. Wiskott, J-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[10] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America*, vol. 14, pp. 1724–1733, 1997.

[11] B. Moghaddam and A. Pentland, "Probabalistic visual recognition for object recognition," *IEEE Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.

[12] Penev P. and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Computation in Neural Systems*, vol. 7, pp. 477–500, 1996.

[13] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings of the Second Conference on Audio- and Video-based Biometric Person Authentication*, Washington, D.C., March 1999.