Face Recognition for Smart Environments

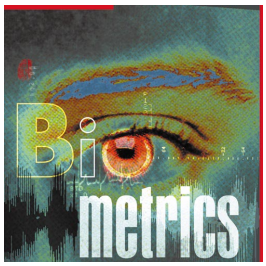
Alex Pentland and Tanzeem Choudhury

Reprint from

IEEE Computer

February 2000

Face Recognition for Smart Environments



Computers of the future will interact with us more like humans. A key element of that interaction will be their ability to recognize our faces and even understand our expressions.

Alex (Sandy) Pentland
Tanzeem Choudhury
MIT Media Laboratory

Smart environments, wearable computers, and ubiquitous computing in general are the coming "fourth generation" of computing and information technology.[1-3] These devices will be everywhere—clothes, home, car, and office—and their economic impact and cultural significance will dwarf those of the first three generations. At the very least, they represent some of the most exciting and economically important research areas in information technology and computer science.

But before this new generation of computers can be widely deployed, those working on interfaces must invent new methods of interaction without a keyboard or mouse. To win wide consumer acceptance, these interactions must be friendly and personalized, which implies that next-generation interfaces will be aware of the people in their immediate environment and, at a minimum, know who they are.

MEANS OF IDENTIFICATION

Given the requirement for determining people's identities, the obvious question is, what technology is best? A wide variety of identification technologies are available, and many have been in widespread commercial use for years. The most common personal verification and identification methods today are password/PIN (personal identification number) systems and token systems (using tokens such as your driver's license). Because such systems are vulnerable to forgery, theft, and lapses in users' memories, biometric identification systems, which use pattern recognition techniques to identify people by their physiological characteristics, are attracting considerable interest. Fingerprints are a classic example of a biometric; newer technologies include retina and iris recognition.

While they are appropriate for bank transactions and entry into secure areas, such biometric technologies have the disadvantage of being intrusive, both physically and socially. They require users to position their bodies 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-grained spatial sensing required. There is an "oracle-like" aspect to the interaction as well. Since people do not recognize each other by such things as retina scans, these types of identification feel intrusive.

The pause-and-declare interaction is useful in high-security applications (the interruption makes people security conscious), but it is exactly the opposite of what is required for a store that recognizes its best customers, an information kiosk that remembers you, or a house that knows the people who live there. Face and voice recognition have a natural place in these next-generation smart environments. They are unobtrusive (they recognize at a distance without a pause-and-declare interaction), they are usually passive (needing no special electromagnetic illumination), they do not restrict user movement, and they are now both low power and inexpensive. Perhaps most important, though, is the fact that humans identify other people by their faces and voices and are likely to be comfortable with systems that use similar means of recognition.

ACHIEVING FACE RECOGNITION

Twenty years ago, the problem of face recognition was considered among the hardest in artificial intelligence and computer vision. Surprisingly, however, over the past decade, a series of successes have made general personal identification appear not only technically feasible but also economically practical.

The apparent tractability of the face recognition problem, combined with the dream of smart environments, has produced a huge surge of interest from funding agencies and from researchers themselves. It

0018-9162/00/$10.00 © 2000 IEEE

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 more than 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, human skin colors lie on a one-dimensional manifold in color space, with color variation primarily due to melanin concentration. Human facial geometry is limited and essentially two-dimensional when people are looking toward the camera. Today, researchers are working on relaxing some constraints of existing face recognition algorithms to achieve robustness under changes due to lighting, aging, rotation in depth, and expression. They are also studying how to deal with variations in appearance due to such things as facial hair, glasses, and makeup—problems that already have partial solutions.

Typical representational framework

The dominant representational approach that has evolved is descriptive rather than generative. This approach uses training images to characterize the range of two-dimensional appearances of objects the system must recognize. Although researchers initially used very simple modeling methods, they now principally characterize appearance by estimating the probability density function (PDF) of the image data for the target class.

For instance, given several examples of a target class Ω—for example, faces—in a low-dimensional representation of the image data, it is straightforward to model the PDF $P(x|\Omega)$ of its image-level features x as a simple parametric function—a mixture of Gaussian distribution functions—thus obtaining a low-dimensional, computationally efficient appearance model for the target class. In other words, we can use example face images to obtain a simple mathematical model of facial appearance in image data.

Once we have learned the PDF of the target class, 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 a system can use 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 current face recognition methods can process video data at 30 frames per second. Several systems can compare an incoming face to a database of thousands of people in under one second—and all on a standard PC!

Dealing with dimensionality

To obtain an appearance-based representation, the image must first be transformed into a low-dimensional coordinate system that preserves the general perceptual quality of the target object's image. This transformation is necessary to address the problem of dimensionality: The raw image data has so many degrees of freedom that it would require millions of examples to directly learn the range of appearances. Typical methods for reducing dimensionality include

- Karhunen-Loève transform (also called principal components analysis),
- Ritz approximation (also called example-based representation),
- Sparse-filter representations (for example, Gabor jets and wavelet transforms),
- Feature histograms, and
- Independent-component analysis.

Face recognition capitalizes on regularities that are peculiar to humans.

These methods all allow efficient characterization of a low-dimensional subspace within the large space of raw image measurements. Once you obtain a low-dimensional representation of the target class—face, eye, or hand—you can use standard statistical parameter estimation methods to learn the range of appearances that the target exhibits in the new, low-dimensional coordinate system. Because of the lower dimensionality, obtaining a useful estimate of either the PDF or the interclass discriminant function requires relatively few examples.

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 by directly modeling the PDF. A simple linear example of such a difference feature is the Fisher discriminant. Systems can also employ discriminant classifiers, such as support vector machines, which attempt to maximize the margin between classes.

FACE RECOGNITION EFFORTS

The subject of face recognition is as old as computer vision because of the topic's practical importance and theoretical interest from cognitive scientists. Despite the fact that other identification methods (such as fingerprints or iris scans) can be more accurate, face recognition has always been a major research focus because it is noninvasive and seems natural and intuitive to users.

Perhaps the most famous early example of a face recognition system is that of Teuvo Kohonen of the Helsinki University of Technology,[4] who demonstrated that a simple neural net could perform face

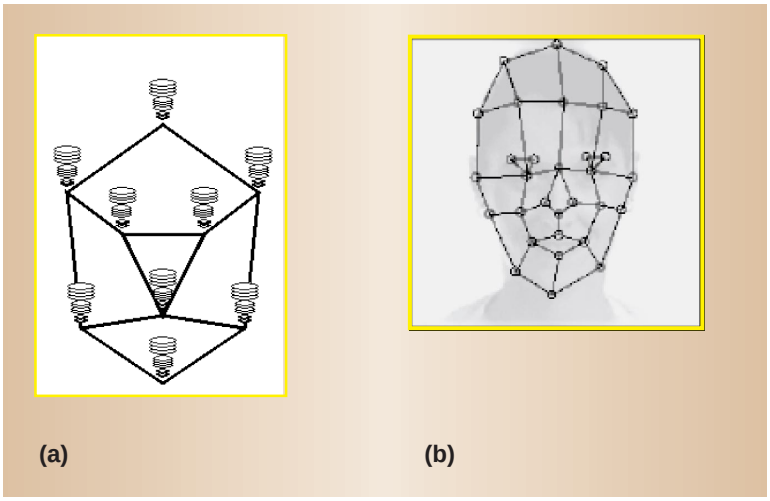


Figure 1. USC's system uses elastic graph matching for face recognition. It creates a face bunch graph from 70 face models to obtain a general representation called (a) an object-adapted grid. The system then (b) matches a given image to the face bunch graph to find the fiducial points. It creates an image graph using elastic graph matching and then compares that image to a database of faces for recognition.

recognition for aligned and normalized images of faces. The network he employed computed a face description by approximating the eigenvectors of the image's autocorrelation matrix. These eigenvectors are now known as *eigenfaces*.

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

Michael Kirby and Lawrence Sirovich of Brown University[5] later introduced an algebraic manipulation that made it easy to directly calculate the eigenfaces. They also showed that it required fewer than 100 eigenfaces to accurately code carefully aligned and normalized face images. Matthew Turk and Alex Pentland of MIT[6] then demonstrated that the residual error when coding with the eigenfaces could be used to detect faces in cluttered natural imagery and to determine the precise location and scale of faces in an image. They then demonstrated that coupling this method for detecting and localizing faces with the eigenface recognition method could achieve reliable, real-time recognition of faces in a minimally constrained environment. This demonstration that a combination of simple, real-time pattern recognition techniques could create a useful system sparked an explosion of interest in face recognition.

Current work

By 1993, researchers claimed that several algorithms provided accurate performance in minimally constrained environments. To better understand the potential of these algorithms, the US Defense Advanced Research Projects Agency and the US Army Research Laboratory established the Feret (face recognition technology) program with the goals of evaluating their performance and encouraging advances in the technology.[7]

Feret identified three algorithms that demonstrated the highest level of recognition accuracy with large databases (1,196 people or more) under double-blind testing conditions: those of the University of Southern California (USC), illustrated in Figure 1;[8] the University of Maryland (UMD);[9] and the Massachusetts Institute of Technology (MIT) Media Laboratory, illustrated in Figure 2.[10] Only two algorithms, those from USC and MIT, are capable of both minimally constrained detection and recognition; the UMD system requires approximate eye locations to operate. Rockefeller University developed a fourth algorithm, illustrated in Figure 3, that was an early contender, but it was withdrawn from testing to form a commercial enterprise.[11] The MIT and USC algorithms have also become the basis for commercial systems.

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

The Feret database testing employs faces with variable positions, scales, and lighting in a manner consistent with mug shot or driver's license photography. With databases of fewer than 200 people and images taken under similar conditions, all four algorithms perform nearly perfectly. Interestingly, even simple correlation matching can sometimes achieve similar accuracy for databases of only 200 people.[7] This is strong evidence that any new algorithm should be tested with databases of at least 200 individuals and should achieve performance over 95 percent on mug-shot-like images to be considered potentially competitive.

In the larger Feret testing (with 1,196 or more images), the performance of the four algorithms are similar enough that it is difficult or impossible to make meaningful distinctions among them (especially if adjustments for date of testing are made). With frontal images taken on the same day, the typical first-choice recognition performance is 95 percent accuracy. For images taken with different cameras and lighting, typical performance drops to 80 percent accuracy. For images taken one year later, the typical accuracy is approximately 50 percent. Still, it should be noted that even 50 percent accuracy is hundreds of times chance performance.

1. The system collects a database of face images.

2. It generates a set of *eigenfaces* by performing principal component analysis (PCA) on the face images. Approximately 100 eigenvectors are enough to code a large database of faces.

3. The system then represents each face image as a linear combination of the eigenfaces.

4. Given a test image, the system approximates it as a combination of eigenfaces. A distance measure indicates the similarity between two images.

1. The system obtains data sets Ω_I and Ω_E by computing intrapersonal differences (matching two views of each individual in the data set) and by computing extrapersonal differences (matching different individuals in the data set).

2. It generates two sets of eigenfaces by performing PCA on each class.

3. The system derives a similarity score between two images by calculating $S = P(\Omega I|\Delta)$, where Δ is the difference between a pair of images. If S is less than 0.5, the system considers the two images to be of the same individual.

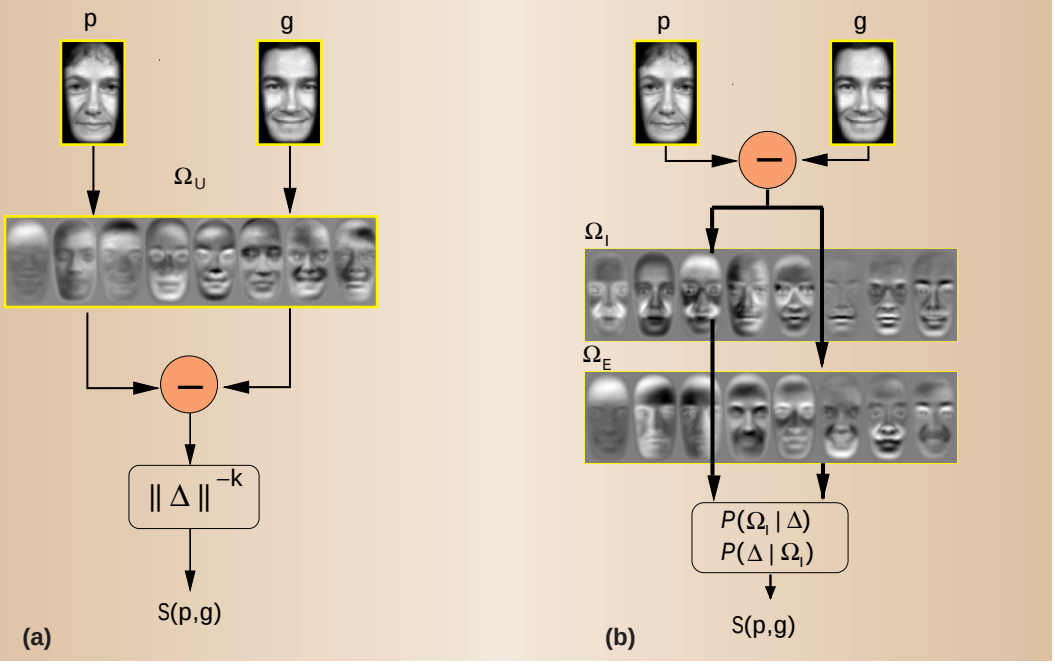


Figure 2. MIT's system of using eigenfaces for face recognition relies on (a) appearance and (b) discriminative models.

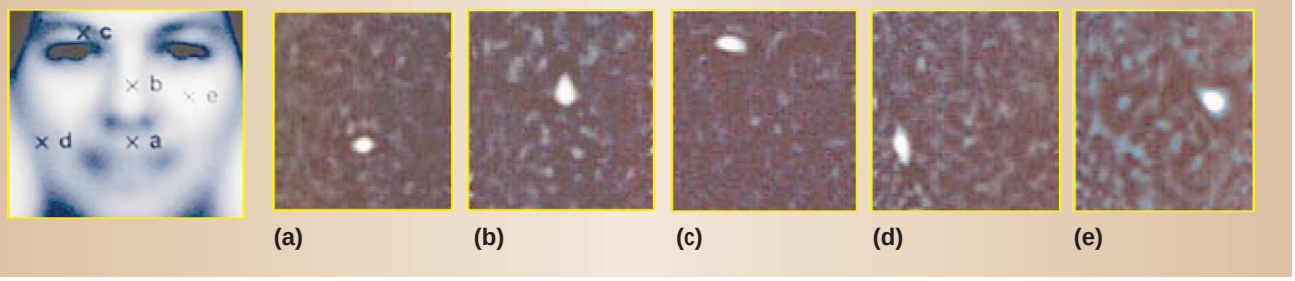


Figure 3. Rockefeller University's system of using local feature analysis for face recognition. The parts marked on the image to the left correspond to receptive fields for the (a) mouth, (b) nose, (c) eyebrow, (d) jawline, and (e) cheekbone. (Reprinted with permission of NYT Pictures)

Growing numbers of applications are using face recognition as the initial step toward interpreting human actions.

Commercial systems and applications

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

- Visionics' FaceIt software is based on the local feature analysis algorithm developed at Rockefeller University. A commercial company in the UK is incorporating FaceIt into a closed-circuit television anticrime system called Mandrake. This system searches for known criminals in video data acquired from 144 closed-circuit camera locations. When a match occurs, the system notifies a security officer in the control room.
- Viisage, another leading face recognition company, uses the eigenface-based recognition algorithm developed at the MIT Media Laboratory. Companies and government agencies in many US states and several developing nations use Viisage's system in conjunction with identification cards—for example, driver's licenses and similar government ID cards.
- Miros uses neural network technology for its TrueFace face recognition software. TrueFace is used by the Mr. Payroll Corp. system in its check cashing system and has been deployed at casinos and similar sites in many US states.

NOVEL APPLICATIONS

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

Face recognition for smart environments

Researchers today are actively building smart environments—visual, audio, and haptic interfaces with environments such as rooms, cars, and office desks.[1,2] In these applications, a key goal is to give machines perceptual abilities that allow them to function naturally with people—to recognize 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 exploring applications for these perceptually aware devices in health care, entertainment, and collaborative work.

Facial-expression recognition interacts with other smart-environment capabilities. For example, a smart system should know whether the user looks impatient—because information is being presented too slowly—or confused because it is coming too fast. Facial expressions provide cues for identifying and distinguishing between these states. Recently, much effort has gone into creating a person-independent expression recognition capability. While there are similarities in expressions across cultures and between people, for anything but the most gross facial expressions, the analysis must account for the person's normal resting facial state—something that definitely isn't the same between people. Consequently, facial-expression research has so far been limited to recognition of a few discrete expressions rather than addressing the entire spectrum of expressions along with their subtle variations. Before a system can achieve a really useful expression analysis capability, it must first be able to recognize and tune its parameters to a specific person.

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.[3] These wearable devices can adapt to a specific user and be more intimately and actively involved in the user's activities. The wearable-computing field 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 aids and context-aware systems. Thus, developers will integrate many future recognition systems with 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 at by whispering it in your ear. The US Army has started testing such devices for use by border guards in Bosnia, and researchers at the University of Rochester's Center for Future Health are looking at them for patients with Alzheimer's disease (see http://wearables.www.media.mit.edu/projects/wearables and http://www.futurehealth.rochester.edu).

FUTURE WORK

Today's face recognition systems work very well under constrained conditions, such as frontal mug shot images and consistent lighting. All current face recognition algorithms fail under the vastly varying conditions in which humans can and must identify other people. Next-generation 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; thus, fusion with other modalities is essential. Technology used in smart environments has to be unobtrusive and allow users to act freely. Wearable systems in particular require the sensing technology to be small, low power, and easily integrable with clothing. Considering all the requirements, systems that use face and voice identification seem to have the most potential for widespread application.

Cameras and microphones today are very small and lightweight, and have been successfully integrated with wearable systems. Audio- and video-based recognition systems have the critical advantage of using the modalities humans use for recognition. Finally, researchers are beginning to demonstrate that unobtrusive audio- and video-based personal identification systems can achieve high recognition rates without requiring the user to be in a highly controlled environment.[12]

The goal of smart environments is to create a space where computers and machines are more like helpful assistants, rather than inanimate objects. Face recognition technology could play a major part in achieving this goal, and it has come a long way in the past 20 years. But to achieve the goal of widespread application in smart environments, next-generation face recognition systems will have to fit naturally within the pattern of normal human interactions and 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. Although substantial research remains to be done, these goals now appear to be within reach. ❖

References

1. M. Weiser, "The Computer for the 21st Century," *Scientific American*, Mar. 1991, pp. 66–76.
2. A. Pentland, "Smart Rooms, Smart Clothes," *Scientific American*, Apr. 1996, pp. 68–76.
3. A. Pentland, "Wearable Intelligence," *Scientific American Presents*, Apr. 1998, pp. 90–95.
4. T. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1989.
5. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *Trans. IEEE Pattern Analysis and Machine Intelligence*, Jan. 1990, pp. 103–108.
6. M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cog. Neuroscience*, Jan. 1991, pp. 71–86.
7. P. Phillips et al., "The Feret Database and Evaluation Procedure for Face Recognition Algorithms," *Image and Vision Computing*, May 1998, pp. 295–306.
8. L. Wiskott et al., "Face Recognition by Elastic Bunch Graph Matching," *Trans. IEEE Pattern Analysis and Machine Intelligence*, July 1997, pp. 775–779.
9. K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *J. Optical Soc. of America*, pp. 1724–1733.
10. B. Moghaddam and A. Pentland, "Probabilistic Visual Recognition for Object Recognition," *Trans. IEEE Pattern Analysis and Machine Intelligence*, July 1997, pp. 696–710.
11. P. Penev and J. Atick, "Local Feature Analysis: A General Statistical Theory for Object Representation," *Network: Computation in Neural Systems*, Mar. 1996, pp. 477–500.
12. T. Choudhury et al., "Multimodal Person Recognition Using Unconstrained Audio and Video," *Proc. 2nd Conf. Audio- and Video-Based Biometric Person Authentication*, Univ. of Maryland, College Park, Md., 1999, pp. 176–181.

Alex (Sandy) Pentland is academic head of the MIT Media Laboratory, Toshiba professor, and codirector of the Center for Future Health. He is interested in the development of new technologies that empower people, including those from developing nations. Pentland has a PhD from MIT. He is a cofounder of the IEEE Face and Gesture Recognition Conference and the IEEE Computer Society Technical Committee on Wearable Information Devices. Contact him at sandy@media.mit.edu.

Tanzeem Choudhury is a graduate student at the MIT Media Laboratory. Her research interests include face recognition, real-time multimodal person identification, and facial expression analysis. Choudhury has an SM from MIT and a BS in electrical engineering from the University of Rochester. Contact her at tanzeem@media.mit.edu.