Looking at People: Sensing for Ubiquitous and Wearable Computing

Alex Pentland, *Member*, *IEEE Computer Society*

Abstract—The research topic of *looking at people,* that is, giving machines the ability to detect, track, and identify people and more generally, to interpret human behavior, has become a central topic in machine vision research. Initially thought to be the research problem that would be hardest to solve, it has proven remarkably tractable and has even spawned several thriving commercial enterprises. The principle driving application for this technology is "fourth generation" embedded computing: "smart"' environments and portable or wearable devices. The key technical goals are to determine the computer's context with respect to nearby humans (e.g., who, what, when, where, and why) so that the computer can act or respond appropriately without detailed instructions. This paper will examine the mathematical tools that have proven successful, provide a taxonomy of the problem domain, and then examine the state-of-the-art. Four areas will receive particular attention: person identification, surveillance/monitoring, 3D methods, and smart rooms/perceptual user interfaces. Finally, the paper will discuss some of the research challenges and opportunities.

Index Terms—Looking at people, face recognition, gesture recognition, visual interface, appearance-based vision, wearable computing, ubiquitious.

1 Introduction

Smart environments, wearable computers, perceptual user interfaces, and ubiquitous computing generally are widely thought to be the coming of "fourth generation" computing and information technology [1], [2], [3]. Because these embedded computing 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.

Although the specifics of such "fourth generation" systems are still far from settled, it is clear that before this new generation of computing can be widely deployed, it must be equipped with sensing technology that allows the computers to be used without detailed instruction and to respond to some situations automatically. Because of the impossibility of having users instruct each of the (possibly dozens) computers these future "smart" environments, the computers must know enough about the people in their environment so that they can act appropriately with the minimum of explicit instruction. Thus, the problem of *context sensing*, which is closely related to the famous *frame problem* of AI,[1] has become a critical problem in the development of fourth generation computing.

1.1 Why Vision Sensors?

Many types of sensor systems are being explored for purposes of context sensing. Bar codes are a well-known solution for connecting physical packages of consumer goods to electronic computer databases [4]. Passive radio-frequency (RF) tags are widely used to prevent theft from stores and new inexpensive versions containing many bits of identification information are becoming a popular way to identify people and perform warehouse inventories [1].

Tagging systems like these are both accurate and cost-effective for identifying packages or cooperating people. This has lend many observers to argue that general-purpose visual object recognition is unnecessary for fourth-generation systems. However all existing tagging systems have some severe limitations: they require close proximity, can provide only limited information (it is hard to imagine a tag that signals when a person smiles), and typically require sending out an electromagnetic probe signal. Thus, while tagging systems make it debatable that machine vision will be a cost-effective solution for recognition of inanimate objects, it is unlikely that tagging systems will be sufficent for interpreting the *human* portion of the computer's context: where people are looking or pointing, what task they are doing, and their affective state (e.g., tired, confused, pleased, stressed, etc.).

Vision systems, therefore, seem to have a natural place in this new generation of computing systems. They are able to track, recognize, and interpret at a distance, are generalizable to novel but similar situations (e.g., one can recognize member of the set of all faces, not just faces that have been trained upon), are usually passive (do not require generating special electromagnetic illumination), and are now both low-power and inexpensive.

It is these general properties, together with fourth-generation system's need to sense humans in the surrounding environment, that has lent a special impetus to the

[1] Roughly, the frame problem is knowing which facts are relevant to the current reasoning problem, and which facts are irrelevant.

• *The author is with the Media Laboratory, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, MA 02139.*
E-mail: pentland@media.mit.edu.

Manuscript received 23 July 1999; accepted 1 Nov. 1999.
Recommended for acceptance by K. Bowyer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110292.

research problems of *looking at people,* that is, to the machine vision problem of detecting, tracking, and identifying people, and more generally to interpreting human behavior.

1.2 Other Research Motivations

Visual sensing of humans is also a central topic for digital libraries. Because people are visual animals, we have begun to collect large databases of digital visual material. It is important to be able to index such databases by their content, and it is interesting to note that the capabilities that are critical for interface applications seem to be exactly those that are most important for indexing digital libraries by their human content.

One can speculate that this is because both interface and database tasks need to extract similar range of semantic content. It appears that when considering the full range of possible application demands, both interface and indexing tasks require a fairly general, substantially complete representation of the humans in the scene, their identities, what they are doing, and their relationships. Because the demands of interface and indexing applications are similarly broad, I will not distinguish between vision tools for perceptual interfaces and those for database indexing.

In addition to practical concerns such as databases or interfaces, another important incentive for investigating these problems is that they are central to understanding human and biological systems. Beyond basic capabilities such as depth perception, the perception of facial identity, expression, and gesture are among the most active areas of psychological research. Machine vision techniques for constructing and utilizing models of facial appearance, expression, and gesture provide important information for computational modeling of biological systems.

1.3 Outline of Paper

This paper will begin by examining the context in which looking-at-people research has arisen and the mathematical tools that have proven successful. The paper will then survey the problem domain, providing a taxonomy of types of applications and types of measurements, and then examine the state of the art. Four areas will receive particular attention: person identification, surveillance/monitoring, 3D methods, and smart rooms/perceptual user interfaces. Finally, the paper will discuss some of the challenges and opportunities facing researchers in this area.

2 History and Framework

It is hard to remember now, but twenty years ago problems like face recognition, person tracking, and gesture recognition were considered among the hardest of machine vision problems and the ones least likely to have quick successes. It was certainly not a problem area that attracted large numbers of researchers or substantial funding. Fortunately, over the last decade there have been a series of successes that have made the general looking-at-people enterprise appear not only technically feasible but economically practical.

The combination of demand from the "fourth generation" systems and the apparent tractability of several aspects of the looking-at-people problem has produced a huge surge of interest from both funding agencies and from vision 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 a database of over 1,000 people [5], commercially available camera systems that perform real-time face tracking for teleconferencing, and companies like IBM, Microsoft, Sony, Siemens, and Mitsubishi are showing simple vision-based gesture recognition interfaces in commercial applications [6], [7], [8], [10], [11] (see Fig. 1).

Many of these early successes came from the combination of well-established pattern recognition techniques with a fairly sophisticated understanding of image generation process. In addition, these methods often capitalized 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), that the unusual spatio-temporal structure of eye blinks that makes them easy to detect, or that human facial geometry is limited and essentially 2D when people are looking toward the camera.

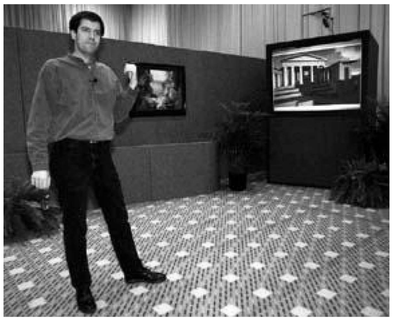
2.1 Representational Framework

The dominant representational approach that has evolved is descriptive rather than generative. Training images are used to characterize the range of 2D appearances of objects to be recognized and for dynamic phenomena (e.g., gestures) to characterize the temporal trajectory of the 2D appearance parameters. This descriptive approach is dominant even in 3D approaches, where texture maps of visible surfaces, 3D shape models, and motion patterns are usually learned from data rather then generated from a priori models.

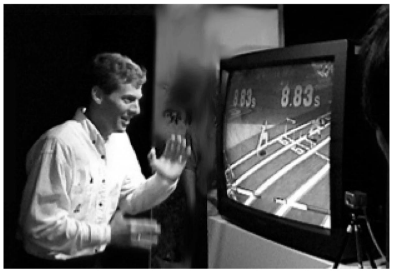
Although initially very simple modeling methods were used, the dominant method of characterizing appearance has fairly quickly became estimation of the probability distribution function (PDF) of the image data for the target class. For instance, given several examples of a target class Ω in a low-dimensional 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 target classes' PDF has been learned, we can use Bayes' rule to perform maximum a posteriori (MAP) detection and recognition. The result is typically a very simple filter-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.

One important variation on this methodology are *discriminative models*, which attempt to model the differences between classes rather than the classes themselves [124]. 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 [13], [14]. One can also employ discriminant classifiers such as Support Vector Machines (SVM) which attempt to maximize the margin between classes [15]. SVMs



(a)



(b)

Fig. 1. (a) The IBM's perceputal interface system (based on the MIT Media Lab pfinder vision system) [6], and (b) the Mitsubishi game playing vision interface that gave rise to the Nintendo camera system [11].

are very closely related to Bayesian methods and Maximum Entropy methods for discriminative learning [16].

2.1.1 *Appearance-Based Methods*

The use of parametric appearance models to characterize the PDF of an object's appearance in the image is related to the idea of a view-based representation [17], [18]. As originally developed, the idea of view-based recognition was to accurately describe the spatial structure of the target object by interpolating between previously seen views. However, most working systems have found that in order to describe natural objects such as faces or hands, which display a wide range of structural and nonrigid variation, it is necessary to extend the notion of "view" to include characterizing the range of geometric and feature variation, as well as the likelihood associated with such variation. Such density-based modeling techniques are now generically known as "appearance-based" methods.

To obtain such 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. The necessity for such a transformation is 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 also employed, including sparse filter representations (e.g., Gabor Jets, Wavelet transforms), feature histograms, independent components analysis, "blobs," and so forth [19], [20], [21], [22], [23], [24], [25], [26].

These methods have in common the property that they allow efficient characterization of a low-dimensional subspace within the overall space of raw image measurements. Once a low-dimensional representation of the target class (face, eye, hand, etc.) has been obtained, then 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 interclass discriminant function.

2.2 Time Domain Phenomena

Appearance-based methods can be applied directly to time domain data by simply characterizing the PDF of the space-time signal. Roughly, one simply considers several frames of image data at once, rather than treating them separately. Such a direct approach has been shown to be useful for many situations, particularly when the space-time target has a fixed duration and viewpoint. Examples are tennis swings seen from the net line or facial movements seen head-on [27], [28].

However, in cases were the evolution in time is complex, due to temporal variability, projective effects, or 3D constraints, such direct modeling of the PDF may be inadequate. Examples of more complex phenomena are words in American Sign Language and pedestrian walking patterns within a plaza. To adequately capture these more complex phenomena, one must allow for a *sequence* of appearance models that are flexibly matched to the image data. The standard mathematical methods for achieving this matching come from control theory, e.g., Kalman filters, or from speech signal processing, e.g., Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs), or from language processing (Dynamic Baysian Networks) [29], [30], [31], [32]. Perhaps, unsurprisingly, these methods are mathematically equivalent in simple applications.

For even longer-term phenomena, e.g., driving behaviors which may take many seconds, sentences in American Sign Language, or coordinated activities like ballroom dancing, the structural richness of HMM or Kalman filter modeling is inadequate. Characterization of these longer-term behaviors requires sequencing of several HMM-like or Kalman-like models together. Methods for accomplishing this range from simple multiple model competition (e.g., multiple-model Kalman filters [33] or sampling methods [34]), to multilevel stochastic models (e.g., multilevel HMMs) [35], [36], to full logic or grammar models that describe complex constraints between lower-level appearance models [37], [38], [39].

3 The Problem Domain

The "looking-at-people" domain is more complex than most areas of image understanding. Like other areas of vision research, it includes geometry, dynamics, complex time evolution, nonrigid objects, etc., and to date, the majority of research papers have dealt with these familiar aspects of the problem. However, the looking-at-people domain also includes *intentionality,* e.g., purposive and communicative behavior, as a central and often unavoidable aspect.

To interpret most gestures, for instance, requires making use of the idea that one person *intends* to communicate some information to another person and that the gesture is *purposefully designed* to accomplish this. For instance, many diectic (pointing) gestures don't actually point *at* the item indicated, only toward it—as seen from the listener's viewpoint. Similarly, facial expressions are typically *communicative acts* not geometric sets: The same quick smile can be ironic in the midst of sad expressions, reassuring in the midst of a instructional conversation, etc.

3.1 A Taxonomy: Channels, Time Scale, Intentionality

We can build a taxonomy of the looking-at-people domain by considering the observation channels, their time scale, and finally their intentionality.

The domain is typically broken down into several *channels* of information: facial, hand, whole-body, and voice. Although these channels can be nearly independently controlled, they typically are used in complementary or redundant manner and must, in general, be considered together. Because voice, gesture, and facial expression are so tightly bound up together, it is unfortunate that the visual and auditory research communities are so separate.

Each channel carries interesting information at a wide range of time scales. At the longest scale are semipermanent *physical attributes* like facial shape and appearance, vocal pitch, timbre, and phonetic peculiarities (e.g., accent), body shape and gait. These long-term characteristics are all useful for identification and are predictive of variables such as age or sex. At shorter time scales are *goal-directed behaviors* which typically have durations ranging from a few seconds to minutes or even hours. Examples are getting out of a car and walking into a building or telling a delivery worker where to place a package. Behaviors are in turn composed of a (multimodal) sequence of individual *actions,* such as pointing, grasping a handle, frowning, etc.[2]

For purposes of analysis, these individual actions are often broken down into "microactions," such as the facial "action units" of the FACS system [42]. However, it is uncertain whether such microactions constitute an important level of representation. People are normally unaware of these microactions and are unable to independently control them. These observations support the argument that microactions may be more a convenient accounting system for psychologists than something intrinsic to the structure of the phenomena.

The final split in our taxonomy is into simple phenomena in which intentionality does not need to be considered and then into behaviors with increasingly complex intentionality. Simple physical observations—the traditional focus of image understanding—typically do not involve intentionality. The shape or appearance of a face, the presence or absence of a person, their position, and body pose are all simple physical observations.

The first point at which intentionality must be considered is the observation of *direct behaviors*. These are behaviors that have only the intention of directly influencing the surrounding physical environment, and include "robotic-like" activities such as direct manipulation, construction, cleaning, etc. To interpret such behaviors it is normally necessary to know about both the person's movements and the objects in the surrounding environment, because the movements' *intended purpose* is to manipulate the object.

In contrast, communicative behaviors have the intention of influencing another agent, something often referred to as higher-order intentionality. Included are most expressions and gestures, even unconscious ones since these have evolved to serve an important role in interpersonal communication. To interpret communicative behaviors, it is normally necessary to know at least something about the context, for instance, are there other people present and what is the goal of the interaction [46], [48].

As an example, consider the gesture of extending an arm and finger together. This can either be a pointing gesture (a communicative behavior), or it can be pushing a button (a direct behavior), or even an unconscious muscle stretch (presumably a nonintentional behavior). It is the presence and relative location of the button or the human observer that differentiates these three behaviors.

To date, most machine vision applications have largely avoided dealing with questions that involve intentionality. This is easy to do if one can limit the context: For instance, if you know that there is no button in the area, but there is a human observer, then extending the arm and finger will probably be a pointing behavior and not a button pushing behavior (but watch out for stretching!). The ability to avoid questions of intentionality is a great advantage for today's applications, but as we move towards more generally competent systems we will have to directly confront the problem of interpreting intentionality.

One area where consideration of intentionality is difficult to avoid is viewpoint. In most vision applications, there are only two viewpoints that must be considered: external and object-centered. These correspond roughly to the third person perspective and the first person perspective. However in the looking-at-people domain there is also a "second person perspective," where the observed person is interacting with you and their intentions *toward you* become a primary consideration.

3.2 Measurements and Features

The above taxonomy tells us something about which measurements and features are likely to be important. Consider, for instance, communicative behaviors and actions. From an evolutionary perspective, these displays are intended to be observed by another human at up to dozens of meters of distance in all sorts of outdoor environments. This requirement, coupled with the psychophysical properties of the human visual system, means that 3D geometry is unlikely to be as important as 2D appearance, that coordinate frames centered on the humans involved are likely to be useful, and that large, high-contrast or relative features are more likely to be significant than small, low-contrast, or absolute measurements. The psy-

[2] Bobick [40] and Quek [41] both suggest related but different temporal decompositions.

chophysics of sign language and similar gesture communication systems bear out this reasoning [43].

In contrast to communicative behaviors, direct behaviors and actions are not intended to be observed by another human, and thus have very different constraints. Since these actions are intended to manipulate 3D objects in the immediate space around the human, it is more likely that 3D properties are important, that object-centered coordinate frames are likely to be employed, and that details of geometry are likely to be significant.

Visual processing of physical characteristics (face recognition, person tracking, etc.) can make use of yet other constraints and heuristics. The fact that people generally look where they are going means that it is easy to acquire frontal face images, and consequently that methods that assume a frontal face image are adequate for most face recognition problems. The fact that people's bodies have fairly standard geometries and typically display a very limited range of poses means that appearance-based methods are useful in detecting, tracking, and pose recognition. The fact that the color of human skin is almost entirely determined by the amount of melanin means that color is a useful cue for detecting and tracking faces and hands. Similarly, the bilateral symmetry and dynamics of the human body mean that simple motion cues can be used to detect and characterize human walking, blinking, etc.

4 The State of the Art

This section will survey the history and current state-of-the-art in the looking at people domain. Rather than an exhaustive survey, however, the focus will be on the early efforts that had the greatest impact on the community (as measured by, e.g., citations), and those current systems have received extensive testing.

In practice, these two criteria often restrict discussion to real-time systems. This is because in the looking-at-people domain we do not have available complete, analytical models, making it is difficult to get a convincing sense of how well a system works unless it has been applied to a large number of people in many different situations.

For more exhaustive surveys, readers are referred to several recent publications. For tracking of the body and hand, see Gavrila [44]. For gesture recognition, see Pavlovic et al. [45]. For linguistic models of gesture, see McNeil [46], Quek [41], or Cassell [48]. For face analysis see Chellappa et al. [49]. For a collection of several historically significant systems and important development efforts, see [50]. For an overview of human modeling techniques see Cerezo et al. [51]. For current research, the reader is referred to the *Proceedings of the IEEE Conferences on Automatic Face and Gesture Recognition* [5], [52], [53].

Current looking-at-people research is largely concentrated in just a few isolated areas, in part because of interest by research sponsors, and in part because certain topics are more accessible to researchers with limited resources. The topic of face recognition is largely motivated by interest in person identification, primarily for unobtrusive access control. Person tracking and simple action recognition (e.g., picking up or dropping off parcels) are motivated primarily by surveillance applications. The topics of facial expression recognition and hand gesture recognition are largely motivated by interest in next-generation "perceptual" interfaces. The topic of 3D face and body tracking, in contrast, is driven at least as much by interest in advanced video coding and 3D image display.

Over time, one can expect to see more integration between these different application areas, as well as between audio and video, with the eventual goal of deriving a reliable semantic analysis from the video data. For review of such multimodal approaches, see [90], [55]. This article, however, will use the currently popular divisions.

4.1 Person Identification via 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 [56]. Despite the fact that other methods of identification (such as RF tags, fingerprints, or iris scans) can be more accurate, face recognition remains of primary interest because of its noninvasive 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 [57], 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, interfeature distances, and other neural-net approaches [58], [59], [61], [60]. 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 [19] 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 [20] then demonstrated that the residual error when coding using the eigenfaces could be used both to detect faces in cluttered natural imagery and that it could also be used for precise determination of the location, scale, and orientation 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 recognition of faces in a minimally constrained environment.

4.1.1 *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 Laboratory established the FERET program with the goals of both evaluating their performance and encouraging advances in the technology [62].

At the time of this writing, there are three algorithms that have demonstrated the highest level of recognition accuracy on large databases (1,196 people or more) under double-blind testing conditions (see Fig. 2). These are the algorithms from University of Southern California (USC) [21], University of Maryland (UMD) [63], and the Massachsetts Institute of Technology (MIT) Media Lab [64]. 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 UMD system requires approximate eye locations to operate. A fourth algorithm was an early contender, developed at Rockefeller University [65], but dropped from testing to form a commercial enterprise. The MIT and USC algorithms have also become the basis for commercial systems.

The MIT and UMD algorithms use a eigenface transform followed by discriminative modeling. The UMD algorithm uses a linear discriminant, while the MIT system employs a quadratic discriminant. The Rockefeller system, although never described in detail, appears to use a sparse version of the eigenface transform, followed by a discriminative neural network. The USC system, 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. 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 mugshot-like images before it can be considered potentially competitive.

In the larger FERET testing (using databases as large as 1,196 people), the performance of these 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 over 95 percent accuracy. For images taken with a different camera and lighting, typical performance drops to 80 to 90 percent accuracy. And for images taken one year later, the typical accuracy is approximately 50 percent. Note that the at-chance accuracy on these tests is typically less than 0.5 percent.

4.2 Surveillance and Monitoring

Security concerns also drive another major area of research: surveillance and monitoring. Work in this area typically uses unobtrusively mounted cameras to detect people, track them, and perform a limited analysis of their behavior. Examples of behavior-related questions are: Did they pick up or drop off any packages? Were they following another person? Were they moving in an atypical manner?

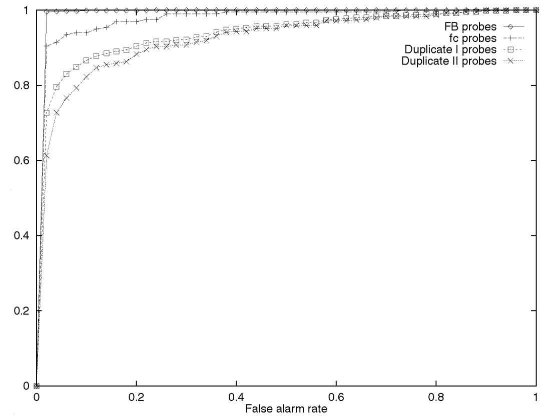


Fig. 2. ROC curves showing best performance for different face recognition tests [62]. The curve marked "FB" is a test with 1,196 people using images taken under similar conditions and at similar times. The other curves are test results for several hundred people under different conditions. The "fc" curve shows accuracy using a different imaging set-up, and the duplicate I and II curves are images taken at progressively later dates (up to a year after the first image).

Some of the first work in this area was by Akita [66], who demonstrated that the human body could be visually tracked, at least in very limited circumstances with strong a priori knowledge. However, throughout the 1980s, robust, real-time person tracking was considered a far-off prospect. Today, however, there are many real-time systems for person detection, tracking, and limited activity classification, and DARPA has sponsored a wide range of research projects in the area [67].

4.2.1 *Current State of the Art*

The focus of most of today's real-time systems is detection and tracking, where an analysis of body pose, gesture, etc., is not required. The main problem addressed by researchers building these systems is reliability of detection and tracking despite shadowing and occlusion, and building a representation of "typical" movement patterns (usually by clustering examples of tracked movements) so that "atypical" movements can be identified. Many use low-level motion analysis [68] to discriminate between different types of action.

CMU's system [69], for instance, extracts moving targets from a real-time video stream, classifies them into predefined categories and then tracks them. The MIT AI Lab's system [70] uses real-time, color-based detection and motion tracking to classify detected objects and learn common patterns of activity. Sarnoff, USC, and Hebrew University's systems [72], [73], [124] use stabilization techniques to detect moving objects from a moving airborne camera. SRI's system [74] uses real-time stereo to detect and track multiple people. Lehigh/Columbia's system [71] uses an omnidirectional camera to detect and track multiple blobs (people, cars, etc.) at the same time.

A special case are systems for detecting human faces; such faces are useful as an input to face recognition systems. Today's systems all function in a similar manner, using a statistical characterization of faces and nonfaces to build a discriminative classifier. This classifier is then used to search over space and scale for image patterns that are likely to be human faces [64], [23], [76], [77].

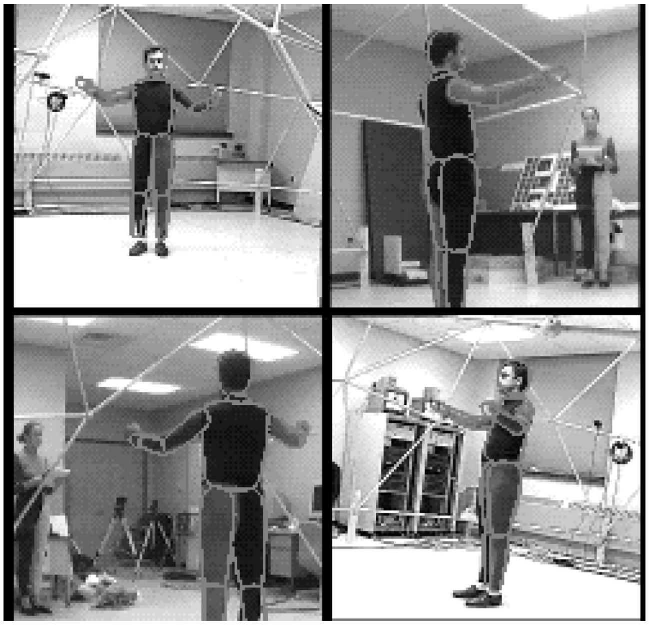


Fig. 3. UMD system for obtaining 3D body model from several image streams [105].

More recently, some surveillance system have begun to perform analysis of both body pose and of the interactions between people and their environment. TI's system [78], for instance, is a general-purpose system for moving object detection and event recognition where moving objects are detected using image change detection, tracked using first-order prediction, and nearest-neighbor matching. Events such as leaving a package or taking an object are recognized by applying predicates to a graph formed by linking corresponding objects in successive frames. The system is limited in that it can not track multiple people.

The MIT Media Lab's systems [79], [80] uses similar methods to detect and track multiple objects, but then employs advanced statistical methods such as coupled HMMs to classify behaviors and interactions (e.g., following, meeting) between agents, and statistical grammars to recognize patterned multiagent behaviors (e.g., being picked up by a car).

The University of Maryland's W4 project [81] uses greyscale statistical difference modeling together with hierarchical shape analysis to track multiple people, including their body parts, in real-time over a wide range of viewing conditions and imaging sensors. It uses statistical matching to classify body pose and can use this information to recover from complex occlusions or to detect carried objects. It uses a variety of motion models to classify an interesting but limited selection of events and interactions between people. (see Fig. 3)

4.3 3D Person Tracking

The third application area, 3D person tracking, is driven as much by the goals of computer graphics and image compression as by the idea that 3D tracking is important in building human-computer interfaces. Imaging standards such as MPEG-4 and MPEG-7 envision accurately tracking people in 3D, allowing recovery of descriptions that are very similar to animation "scripts" in computer graphics. Such high-level descriptions are envisioned as being useful both for very-low-bit-rate image transmission and for searching video databases.

In addition some researchers believe that an intrinsically 3D representation of human movement (as opposed to an appearance-based description of 3D movement) will produce better recognition of gesture, expression, and action. As argued above, however, there is substantial reason to believe that this will not be the case, and current 3D machine vision systems do not perform substantially better than appearance-based systems.

Early researchers in this area were Badler and Smoliar [83] and Hogg [82], who proposed a model-driven, analysis-synthesis approach in an attempt to track people for the purpose of driving computer graphics animations and Mase [107] who used facial line features to estimate 3D head position in real-time. Interestingly, these two basic approaches—analysis-synthesis using an articulated (or flexible) body model, and estimation from feature tracking—remain the dominant methods in this area, although they are almost always combined with more typical appearance-based methods for the actual image matching step.

The 3D tracking problem is difficult because of the nonrigid and articulated nature of the human body, lack of reliable features, frequent occlusions, and loose clothing. This has lead most researchers to employ multiple cameras and to estimate only some parameters (e.g., skeleton joint angles) rather than attempt to recover a full volumetric model. Interestingly, some researchers have taken an approach driven purely by computer graphics requirements: they attempt to recover a volumetric model using dozens of cameras, but make no attempt to segment the data into body parts [85], [86], [87].

4.3.1 *Current State of the Art*

Today there are many systems that can produce an estimate of 3D head pose in real time. Most track facial features in 2D and then use triangulation methods to derive 3D estimates [88], [89], [27], [90], [91]. Several systems also track eye features, allowing an estimate of gaze [90], [88].

A few systems can track both head pose and face shape in real time. The University of Pennsylvania's 3D head and face tracking system uses a finite-elements method (FEM) model running on special hardware to track facial deformation in real time [92]. The MIT Media Lab's system uses point features and a statistical model of the range of 3D head shapes in an Extended Kalman Filter (EKF) formulation to track head pose and shape in real-time with a single camera and single computer [93]. The pose-tracking portion of the MIT system is available commercially in the Alias/Wavefront MayaLive system [94]. Several other systems [98], [97], [95], [96], [99] have demonstrated the capability to track facial pose and shape at extremely fine scale, but are not real-time and thus have not yet been applied to a large and diverse range of data.

Systems are now appearing that can also perform real-time tracking of the human body in 3D. The University of Pennsylvania system uses a FEM formulation running on special-purpose processors and three orthogonal cameras to extract 3D joint angles and an estimate of body shape [100].

The MIT Media Lab DynaMan system uses a fully dynamic skeleton model driven by low-level visual tracking of face and hand "blobs"; the system runs on a PC and uses two cameras [103]. This system is in regular use at several sites around the world. The CalTech arm-tracking system [102] tracks upper and lower arm in real-time using a single camera and a standard computer. The CMU DigitEyes system uses a single camera and special purpose hardware to track hand pose, including 27 joint angles, in real time [101]. Several other systems [104], [95], [105] have demonstrated that they are capable of 3D tracking articulated motion, but are not real-time and thus have not yet been applied to a large and diverse range of data. (see Fig. 3).

4.4 Smart Rooms and Perceptual Interfaces

The construction of human-friendly, next-generation interfaces is likely to be the single most widespread application of future looking-at-people technology. It is well-known that people normally communicate using a subtle combination of gesture, facial expression, body language, and vocal prosody in conjunction with spoken words [46]. In order to obtain interfaces that are qualitatively better than what we have today, it seems clear that machines must be able understand and generate at least some of these same communicative elements [47], [48], [41]. Most recently, a wide variety of industrial leaders, including Gates of Microsoft [108], have declared development of visual interfaces to be a high priority.

Such perceptual user interfaces must have a variety of elements. They must of course be able to detect, recognize, and track people, in common with security and coding applications, but they must also be able to recognize facial expressions and hand gestures, and then integrate these cues with audio processing such as pitch tracking and word recognition.

Building such an integrated perceptual interface requires substantial programming, equipment, and organizational resources. Consequently, most university researchers tend to focus on recognition of individual cues such as hand gestures or facial expressions, and there are now dozens of real-time systems capable of robustly recognizing simple hand or head gestures, and simple facial expressions.

These systems are now capable of very accurate recognition of several dozen hand gestures [45] and recognition of a few facial expressions with useful accuracy [49]. However, rather than focusing on these individual cues, this section will instead describe efforts to produce a more integrated interface. Readers are referred to [45] for a recent survey of gesture recognition systems, and to [49] for a survey of face analysis systems. In addition, the Proceedings of the *IEEE Conferences on Automatic Face and Gesture Recognition* [52], [5], [53] are perhaps the best source for most recent developments in analysis of these component cues.

Perhaps the first effort at building a perceptually aware "Smart Room" was Kruger's early 1980s "Video Place" demonstrations [106]. These systems used special-purpose video hardware and special illumination/viewing techniques to find and track user's heads and hands in real-time, and allowing them to manipulate images, video, etc., unencumbered by interface devices. Krueger's systems, although limited in capability, had an immense impact on the technical community's imagination and brought the term "virtual reality" into common use. Such systems have since become a standard element in video games and film production.

Following Kruger's demonstrations and as part of the general dissatisfaction with keyboard, desktop-like graphical user interfaces, and mouse, researchers began to develop perceptually-aware environments. Perhaps the first attempts at building such integrated, perceptually-aware environments were at NTT [107], Siemens [10], and the MIT Media Lab [2].

NTT's system, which began with a "headreader" in 1987 [84] and was gradually extended to include hand gestures, body tracking, and facial expressions throughout the early 1990s [107]. The system used contour geometry to track the head in 3D and used contours coupled with special illumination to track head and hands. Facial features (which were not tracked in real time) were interpreted from optical flow.

The Siemen's GestureComputer project [10] began in 1991 and used color cues and background comparison to track the head and hands in real-time, including partial estimates of 3D orientation. The system could also interpret several hand gestures. Applications include Command and Control in VR environments, Smart Camera Teleconferencing, and interfaces for kiosks and desk. The system has been tested on a large number of users.

The MIT Media Lab's Smart Rooms project [2] began in 1989 with multiple-person tracking of face and body using region-based statistical methods; the first systems also included face and simple body pose recognition [20]. The MIT system was extended during the early 1990s to include real-time hand gesture analysis, real-time face expression analysis, and joint audio-video analysis [9]. Applications include games, command and control in VR environments, and 3D information browsing. The system has been tested on a large number of users at many sites.

4.4.1 Current State of the Art

In the last half of the 1990s, these early efforts were joined by a host of other projects at places such as University of Illinois [109], ATR [107], Osaka University [88], [111], Georgia Tech [112], Compaq (formerly DEC) Cambridge Research Labs [118], the MIT AI Lab [113], IMAG in Grenoble, France [114], Microsoft [119], the University of Maryland [120], Xerox PARC [110], the MIT Media Lab's Smart Desk [116] and KidsRoom [117] projects, and Interval Research [115] to name just a few of the best-known. The range of applications has also grown, to include scientific visualization, augmenting information flow in workgroups, home safety and eldercare, efficient control of building services, along with early applications such as games, VR, and information browsing. Unfortunately, it is difficult to summarize the current state-of-the-art, both because of the complexity of the systems and because of variation in applications. Perhaps the best approach is to summarize the component capabilities and limitations.

There are several systems that track people in 3D, and can recognize up to a few dozen hand/body gestures reliably. Many of these systems can also track several people at once. Several systems that focus on the desktop can track a few simple facial gestures reliably, produce coarse estimates of gaze, and recognize up to a few dozen hand gestures. Several systems are also beginning to integrate audio and visual processing to produce a more natural, unified interface (see Fig. 4).

However, current systems are only beginning to do both fine-scale analysis (e.g., user facial expression) and large-scale analysis (e.g., user body pose). Similarly, most systems that can track multiple people cannot estimate body pose in 3D. The primary problem seems to be a combination of limited camera resolution, computer power, and viewpoint; in order to accomplish both fine-scale and broad-area analysis therefore seems to require accurate, real-time integration of multiple cameras.

Current systems also do not analyze most of the "normal" gestures that people spontaneously generate. This means that communication between the user and computer is restricted to a few "special" gestures. This limitation is largely because the structure of these spontaneous gestures remains poorly understood and because interpretation of such gestures seems to require a deeper understanding of language and user's intention.

Finally, few systems integrate audio and video information except at the simplest level. Most run speech recognition software in parallel with gesture recognition and combine these two independent information streams in the simplest possible manner. Interactions between the semantic content of the speech and the gesture are ignored, along with nonlinguistic cues such as facial expressions and speech prosody.

5 Challenges and Opportunities

The above list of current perceptual interface limitations is a good place to begin discussion of the challenges and opportunities that face looking-at-people research. In addition to the problems of integrating multiple cameras, better understanding of human communicative behavior, and deeper integration of audio with video, there are a number of related issues.

Perhaps the issue that comes most easily to mind—-because it is familiar to readers of *IEEE Pattern Analysis and Machine Intelligence*—is interpretation in 3D. As discussed previously, many situations do not require 3D interpretation. However, robust performance of face recognition and interpretation, accurate interpretation of diectic gestures, and interpretation of manipulation motions are all likely to benefit from more accurate 3D tracking of the human body. We have seen that significant progress is being made on real-time 3D tracking of the human body, so we can expect that these 3D methods will soon be applied to the whole range of looking-at-people interpretation problems.

Another issue that is typical of all visual processing is that of occlusion and resolution. In the looking-at-people domain, this problem is most acute when looking at crowds of people: Typically, only portions of each person are visible and then often at very low resolution. The problem is in general intractable, however interesting progress is being made using statistical methods, which essentially try to guess body pose, etc., from whatever image information is available. Perhaps the most promising practical method for addresssing this problem is through the use of multiple cameras. While camera selection and fusion are significant problems in their own right, the availability of information from several cameras can be extremely helpful.

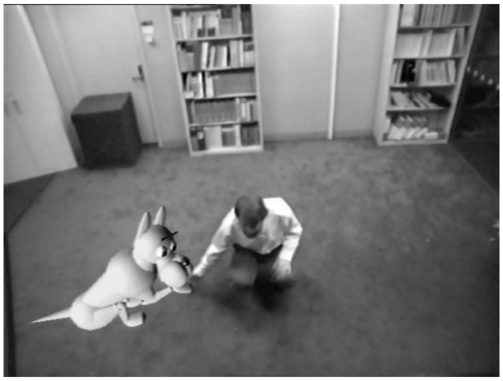


Fig. 4. The ALIVE system used cameras and microphones to interact with virtual creatures [9].

5.1 Audio Input

Audio interpretation of people is at least as important as visual interpretation. Although much work as been done on speech understanding, virtually all of this work assumes a closely-placed microphone for input and a fixed listening position. Speech recognition applications, for instance, typically require near-field $(< 1.5m)$ microphone placement for acceptable performance. Beyond this distance the signal-to-noise ratio of the incoming speech affects the performance significantly; most commercial speech-recognition packages typically break down over a 4 to 6 dB range.

The constraint of near-field microphone placement makes audio interpretation very difficult in an unconstrained environment, so it is necessary to find a solution that allows the user to move around with minimal degradation in performance. One possible solution to this problem involves phased-array microphones, which can "track" users as they move around. However, audio-only solutions can not track silent people or people in very noisy environments so researchers are beginning to investigate whether combined audio and video processing can provide better solution.

Even after acquisition of clean audio input, there is the problem of how to interpret the combined audio-visual signal. Current systems almost universally interpret audio and visual input separately, combining them only at the end [54], [55]. Yet this is almost certainly incorrect; humans show very significant, complex interactions between audio and visual inputs [48]. Because of the complexity of the phenomena and the general lack researchers with expertise in both domains, understanding joint audio-visual interpretation poses a significant research challenge.

5.2 Behavior Understanding

One of the most interesting challenges facing the looking-at-people area is that of better understanding of human behavior. Progress to date has been made largely by focusing on problems that can be solved independent of context. Face recognition, surveillance, person tracking, and recognition of standardized gestures are all examples of such problems. However to be generally useful systems must be able to adjust for individual differences, become more sensitive to task and environmental constraints, and be able to relate face and hand gestures to the semantics of the human-machine or human-human interaction.

Indeed, a key idea of perceptual interfaces is that they must be adaptive both to overall situation and to the individual user. As a consequence, many research groups are beginning to focus on learning user behaviors and how they vary as a function of the situation. For instance, studies have shown that a significant problem in designing effective interfaces is the difficulty in anticipating a person's word choice and associated intent: even in *very* constrained situations, different people choose different words and gestures to mean exactly the same thing [121]. The result is that most current multimodal interfaces are difficult to learn, and often feel unnatural.

One source of help for these problems is *machine learning*: rather than having a priori rules, we can potentially learn the rules by watching the user. For instance, in recent years there has been very significant progress by using machine learning tools to construct statistical models of human behavior. By combining observations of human behavior together with specific classes of graphical models, researchers have been able to derive statistical *rules* of human behavior. These rules have then been used to create context-sensitive and personalized interactions within a variety of interface tasks.

Such learning is particularly important in personal environments. People don't want a "smart home" that works like everyone else's, they want their home to follow their own patterns and peculiarities. Such personalization is even more important with wearable devices: because small size makes the interface so limited, the device has to "do the right thing" with only limited instruction. This seems possible only if the wearable device has learned your habits and preferences.

5.3 First Person Perspective

So far the discussion has centered mostly on sensor interpretation from the third and second person perspective, where the cameras and microphones are either watching people's behavior as they roam about a space or watching people's behavior as they interact with the computer's interface. However, when we build the 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], [123], [122]. Such devices can be more intimately and actively involved in the user's activities.

Although most early wearable devices have relied on sensors like GPS, sonar, or accelerometers for context sensing, it is likely that vision sensors will play a special role because of their ability to sense humans and their movements. 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 their name in your ear. Perhaps more interestingly, a camera mounted in a baseball cap (Fig. 5) can observe the user's hands and feet. This could allow a "computer assistant" to observe the user's gestures and body motion, reducing the need to tell the computer what you are doing. Such a camera can act as an interface for the computer. For example, wearable hand tracking can be used for recognizing American Sign Language [31] (see Fig. 5).

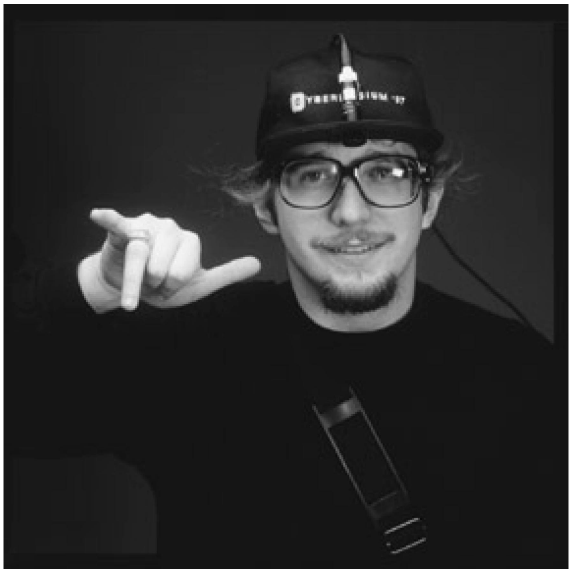


Fig. 5. A nearly-invisible head mounted display (HMD) and a hat-mounted camera to observe the user's hand gestures.

The techniques and problems that confront us in third person observation of people seem to be rather different than those found when using a person-mounted camera. Not only do people move more quickly and erratically than most robotic cameras, but the applications seem different as well. Augmented reality interfaces using a head-mounted display, for instance, are seen as a very important application of head-mounted camera systems, while there is no obvious corresponding task in systems with a third person perspective.

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.

5.4 Privacy

The general goal of looking at people research is to make machines that are aware of the people that interact with them. The idea is that machines should know who we are, see our expressions and gestures, and hear the tone and emphasis of our voice. However, when such perceptually-aware machines are tightly networked together, as in proposals for ubiquitous or pervasive computing environments, we obtain a capacity to concentrate information about people that closely resembles George Orwell's dark vision of a government that can monitor and control your every move.

This is a very serious issue, one that could lead, at an extreme, to society outlawing computers with cameras except in a few special applications. Many experts in the area of privacy believe that unless the computer vision community takes these privacy issues seriously, it risks having its work restricted by law.

However, it is important to note that it is not the cameras that are the problem here, but the networking. Who cares if your house's door knows you came home at 2a.m., unless it can tell your neighbor? This suggests several methods of addressing the privacy problem. One particularly simple approach is simply to avoid the concentrating information except under limited, controlled circumstances. The argument motivating this "quasilocal" approach is that local perceptual intelligence, combined with relatively sparse, user-initiated networking, can provide most of the benefits of ubiquitous networked computing, while at the same time, making it more difficult for outsiders to track and analyze people's behavior.

6 Conclusion

It is now possible to track people's motion, identify them by facial appearance, and recognize simple actions in real time using only modest computational resources. By using this perceptual information, researchers have been able to build smart, perceptually aware environments that can recognize people, understand their speech, allow them to control computer displays without wires or keyboards, communicate by sign language, and warn them they are about to make a mistake.

Researchers are now beginning to apply such perceptual intelligence to a much wider variety of situations. For instance, there are prototypes of displays that know if you are watching, credit cards that recognize their owners, chairs that adjust to keep you awake and comfortable, and homes that know what the kids are doing. Extrapolating these trends it is now possible to imagine environments where the distinction between inanimate and animate objects begins to blur and the objects that surround us become more like helpful assistants or playful pets than insensible tools.

References

[1] M. Weiser, "The Computer for the 21st Century," *Scientific Am.*, vol. 265, no. 3, pp. 66-76, Sept. 1991.

[2] A. Pentland, "Smart Rooms, Smart Clothes," *Scientific Am.*, vol. 274, no. 4, pp. 68-76, 1996.

[3] A. Pentland, "Wearable Intelligence," *Scientific Am. Presents,* vol. 9, no. 4, pp. 90-95, 1998.

[4] R. Stein, S. Ferrero, M. Hetfield, A. Quinn, and M. Krichever, "Development of a Commerically Successful Wearable Data Collection System," *Proc. IEEE Int'l Symp. Wearable Computers,* pp. 18-24, Pittsburgh, Oct. 1998.

[5] *Proc. IEEE Int'l Conf. Face and Gesture Recognition,* I. Essa, ed., Killington, Vt., IEEE CS Press, Oct. 1996.

[6] M. Lucente, G.-J. Zwart, and A. George, "Visualization Space: A Testbed for Deviceless Multimodal User Interface, Intelligent Environments," *Proc. AAAI Spring Symp. Series,* pp. 87-92, Stanford Univ., Mar. 1998.

[7] M. Turk, "Visual Interaction with Lifelike Characters," *Proc. IEEE Int'l Conf. Face and Gesture Recognition,* pp. 368-373, Killington, Vt., Oct. 1996.

[8] J. Rekimoto, Y. Ayatsuka, and K. Hayashi, "Augment-Able Reality: Situated Communication through Physical and Digital Spaces," *Proc. IEEE Int'l Symp. Wearable Computers,* pp. 18-24, Pittsburgh, Oct. 1998.

[9] P. Maes, B. Blumburg, T. Darrell, and A. Pentland, "ALIVE: An Artificial Life Interactive Environment," *Proc. SIGGRAPH '93—Visual,* pp. 115, 1993.

[10] C. Maggioni and B. Kammerer, *GestureComputer: History, Design, and Applications, in Computer Vision for Human-Machine Interaction.* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[11] W. Freeman and C. Weissman, "Television Control by Hand Gestures," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 179-183, Zurich, Switzerland, June 1995.

[12] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabalistic Matching for Face Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 30-35, Nara, Japan, Apr. 1998.

[13] D. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.

[14] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class-Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[15] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer-Verlag, 1995.

[16] T. Jaakkola, M. Meila, and T. Jebara, "Maximum Entropy Discrimination," *Proc. Conf. Neural Information Processing,* Denver, Dec. 1999.

[17] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, pp. 992-1,006, 1991.

[18] T. Poggio and S. Edelman, "A Network that Learns to Recognize Three-Dimensional Objects," *Nature,* vol. 343, pp. 263-266, 1990.

[19] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 103-108, Jan. 1990.

[20] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[21] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 775-779, July 1997.

[22] R. Rao and D. Ballard, "An Active Vision Architecture Based on Iconic Representations," *Artificial Intelligence,* vol. 78, pp. 461-505, 1995.

[23] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian Detection Using Wavelet Templates," *Trans. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 193-199, 1997.

[24] B. Schiele and J. Crowley, "Probabilistic Object Recognition Using Multidimensional Receptive Field Histograms," *Proc. 13th Int'l Conf. Pattern Recognition,* vol. B, pp. 50-54, 1996.

[25] J. Bell and T. Sejnowski, "An Information Maximisation Approach to Blind Separation and Blind Deconvolution," *Neural Computation,* vol. 7, no. 6, pp. 1,129-1,159, 1995.

[26] R. Kauth, A. Pentland, and G. Thomas, "BLOB: An Unsupervised Clustering Approach to Spatial Preprocessing of MSS Imagery," *Proc. 11th Int'l Symp. Remote Sensing of the Environment,* Center for Remote Sensing Information, Ann Arbor, Mich., Apr. 1977.

[27] Y. Yacoob and L. Davis, "Computing Spatio-Temporal Representations of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 6, pp. 636-642, June 1996.

[28] K. Mase and A. Pentland, "Lip Reading: Automatic Visual Recognition of Spoken Words," *Proc. Opt. Soc. Am. Topical Meeting on Machine Vision,* pp. 1,565-1,570, Cape Cod, Mass., June 1989.

[29] T. Darrell, I. Essa, and A. Pentland, "Task-Specific Gesture Analysis in Real-Time Using Interpolated Views," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 12, pp. 1,236-1,242, Dec. 1996.

[30] J. Yamamoto, J. Ohya, and K. Ishii, "Recognizing Human Actions in Time-Sequential Images Using Hidden Markov Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 379-385, 1992.

[31] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition using Desk and Wearable Computer Based Video," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 12, pp. 1,371-1,375, Dec. 1995.

[32] V. Pavlovic, B. Frey, and T. Huang, "Classification Using Mixed-State Dynamic Bayesian Networks," *Proc. IEEE Computer Vision and Pattern Recognition,* vol. 2, pp. 609-615, 1999.

[33] A. Willsky, "Detection of Abrupt Changes in Dynamic Systems, Detection of Abrupt Changes in Signals and Dynamical Systems," *Lecture Notes Control and Information Sciences.* no. 77, Basseville and Benvieniste, eds., Springer-Verlag, 1986.

[34] A. Blake, M. Isard, and D. Reynard, "Learning to Track the Visual Motion of Contours," *Artificial Intelligence,* vol. 78, pp. 101-134, 1995.

[35] M. Friedmann, T. Starner, and A. Pentland, "Synchronization in Virtual Realities," *Presence,* vol. 1, no. 1, pp. 139-144, 1992.

[36] A. Pentland and A. Liu, "Modeling and Prediction of Human Behavior," *Neural Computation,* vol. 11, pp. 229-242, 1999.

[37] H.H. Nagel, H. Kollnig, M. Haag, and H. Damm, "Association of Situation Graphs with Temporal Variations in Image Sequences," *Proc. European Conf. Computer Vision,* vol. 2, pp. 338-347, 1994.

[38] Y. Kuniyoshi and H. Inoue, "Qualitative Recognition of Ongoing Human Action Sequences," *Proc. Int'l Joint Conf. on Artifical Intelligence,* pp. 1,600-1,609. 1993.

[39] J. Siskind, "Grounding Language in Perception," *Artificial Intelligence Rev.,* vol. 8, pp. 371-391, 1994.

[40] A. Bobick, "Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion," *Proc. Royal Soc. B.,* special issue knowledge-based vision in man and machine,vol. 352, pp. 1,270-1,281, 1997.

[41] F. Quek, "Eyes in the Interface," *Image and Vision Computing,* vol. 13, 1995.

[42] P. Ekman and W. Friesen, *Facial Action Coding System.* Palo Alto, Calif.: Consulting Psychologist Press, 1978.

[43] G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible Encoding of ASL Image Sequences at Extremely Low Information Rates," *Computer Vision, Graphics, and Image Processing,* vol. 31, pp. 335-391, 1985.

[44] D. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding,* vol. 73, no. 1, pp. 82-98, 1998.

[45] V. Pavlovic, R. Sharma, and T Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 677-695, July 1997.

[46] D. McNeil, *Hand and Mind: What Gestures Reveal About Thought.* Univ. of Chicago Press, 1992.

[47] R. Bolt, "Put-That-There: Voice and Gesture at the Graphics Interface," *Computer Graphics,* vol. 14, no. 3, pp. 262-270, 1980.

[48] J. Cassell, "A Framework for Gesture Generation and Interpretation," *Computer Vision for Human-Machine Interaction.* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[49] R. Chellappa, C. Wilson, and S. Sirohev, "Human and Machine Recognition of Faces: A Survey," *Proc. IEEE,* vol. 83, no. 5, pp. 705-740, 1995.

[50] *Computer Vision for Human-Machine Interaction.* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[51] E. Cerezo, A. Pina, and F. Seron, "Motion and Behavior Modeling: State of Art and New Trends," *The Visual Computer,* vol. 15, pp. 124-146, 1999.

[52] *Int'l Conf. Automatic Face and Gesture Recognition,* Zurich, Switzerland, June 1995.

[53] *Proc. IEEE Conf. Automatic Face and Gesture Recognition,* Nara, Japan, Apr. 1998.

[54] A. Waibel, M. Vo, P. Duchnowski, and S. Manke, "Multimodal Interfaces," *Artificial Intelligence Rev.,* vol. 10, pp. 299-319, 1995.

[55] R. Sharma, V. Pavlovic, and T. Huang, "Toward Multimodal Human-Computer Interface," *Proc. IEEE,* vol. 86, no. 5, pp. 853-869, 1998.

[56] D. Valentin, H. Abdi, A. O'Toole, and G. Cottrell, "Connectionist Models of Face Processing: A Survey," *Pattern Recognition,* vol. 27, pp. 1,208-1,230, 1994.

[57] T. Kohonen, *Self-Organization and Associative Memory,* Berlin: Springer-Verlag, 1989.

[58] T. Kanade, "Computer Recognition of Human Faces," *Interdisciplinary Systems Res.,* vol. 47, 1977.

[59] I. Craw, H. Ellis, J.R. Lishman, "Automatic Extraction of Face Features," *Pattern Recognition Letters,* vol. 5, pp. 183-187, 1987.

[60] H. Abdi, "Generalized Approaches for Connectionist Auto-Associative Memories: Interpretation, Implication, and Illustration for Face Processing," *Artifical Intelligence and Cognitive Science,* pp. 151-164, 1988.

[61] G.W. Cottrell and M.K. Fleming, "Face Recognition Using Unsupervised Feature Extraction," *Proc. Int'l Neural Network Conf.,* pp. 322-325, 1990.

[62] 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.

[63] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *J. Optical Soc. Am. A.,* vol. 14, pp. 1,724-1,733, 1997.

[64] B. Moghaddam and A. Pentland, "Probabalistic Visual Recognition for Object Recognition," *IEEE Trans. Pattern Anaylsis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[65] P. Penev and J. Atick, "Local Feature Analysis: A General Statistical Theory for Object Representation," *Network: Computation in Neural Systems,* vol. 7, pp. 477-500, 1996.

[66] K. Akita, "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition,* vol. 17, no. 4, pp. 73-83, 1984.

[67] *Proc. DARPA Image Understanding Workshop,* Monterey Calif. San Francisco: Morgan Kaufmann, Nov. 1998.

[68] R. Polana and R. Nelson, "Recognizing Activities," *Proc. IEEE Int'l Conf. Computer Vision,* 1994.

[69] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving Target Detection and Classification from Real-Time Video," *Proc. IEEE Workshop Applications of Computer Vision,* 1998.

[70] E. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using Adaptive Tracking to Classify and Monitor Activities in a Site," *Proc. IEEE Conf. Computer Vision and Pattern Recogition,* pp. 22-29, 1998.

[71] T. Boult, "Frame-Rate Multi-Body Tracking for Surveillance," *DARPA Image Understanding Workshop,* Monterey, Calif. San Francisco: Morgan Kaufmann, Nov. 1998.

[72] A. Selinger and L. Wixson, "Classifying Moving Objects as Rigid or Non-Rigid Without Correspondences," *DARPA Image Understanding Workshop,* Monterey, Calif., San Francisco: Morgan Kaufmann, Nov. 1998.

[73] F. Bremond and G. Medioni, "Scenario Recognition in Airborne Video Imagery," *Proc. DARPA Image Understanding Workshop,* Monterey, Calif. San Francisco: Morgan Kaufmann, Nov. 1998.

[74] K. Konolige, Small Vision Systems: Hardware and Implementation *DARPA Image Understanding Workshop,* Monterey, Calif., San Francisco: Morgan Kaufmann, Nov. 1998.

[75] K.K. Sung and T. Poggio, "Example-Based Learning for View-Based Face Detction," *Proc. DARPA Image Understanding Workshop,* vol. II, pp. 843-850, Monterey, Calif., San Francisco: Morgan Kaufmann, Nov. 1994.

[76] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 203-208, 1996.

[77] H. Schneiderman and T. Kanade, "Probabalistic Modeling of Local Appearance and Spatial Relationships for Object Recognition," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 45-51, 1998.

[78] T. Olson and F. Brill, "Moving Object Detection and Event Recognition Algorithms for Smart Cameras," *Proc. DARPA Image Understanding Workshop,* pp. 159-175, Monterey, Calif. San Francisco: Morgan Kaufmann, 1997.

[79] N. Oliver, B. Rosario, and A. Pentland, "Statistical Modeling of Human Interactions," *Proc. IEEE Conf. Computer Vision and Pattern Recogntion,* 1998.

[80] Y. Ivanov, C. Stauffer, B. Bobick, and W.E.L. Grimson, "Video Surveillance of Interactions," *Proc. IEEE Workshop Video Surveillance,* Fort Collins, Colo., June 1999.

[81] I. Haritaoglu, D. Harwood, and H. Davis, *W4: Who, What, When, Where: A Real-Time System for Detecting and Tracking People.* 1998.

[82] D. Hogg, "Model-Based Vision: A program to See a Walking Person," *Image Vision Computing,* vol. 1, no. 1, pp. 5-20, 1983.

[83] N. Badler and S. Smoliar, "Digital Representations of Human Movement," *ACM Computing Surveys,* vol. 11, no. 1, pp. 19-38, 1979.

[84] K. Mase, Y. Suenaga, and T. Akimoto, "Head Reader: A Head Motion Understanding System for Better Man-Machine Interaction," *Proc. IEEE Systems, Man, and Cybernetics,* pp. 970-974, Nov. 1987.

[85] P. Narayanan, P. Rander, and T. Kanade, "Constructing Virtual Worlds Using Dense Stereo Processing," *Proc. Int'l Conf. Computer Vision,* Greece, 1998.

[86] K. Kutulakos and C. Dyer, "Calibration-Free Augmented Reality," *IEEE Trans. Visualization and Computer Graphics,* vol. 4, no. 1, pp. 1-20, Jan.-Apr. 1998.

[87] S. Moeszzi, A. Katkere, D. Kuramura, and R. Jain, "Reality Modeling and Visualization from Multiple Video Sequences," *IEEE Computer Graphics and Applications,* vol. 16, no. 6, pp. 58-63, 1996.

[88] M. Yachida and Y. Iwai, "Looking at Human Gestures," *Computer Vision for Human-Machine Interaction.* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[89] N. Oliver, F. Berard, J. Coutaz, and A. Pentland, "LAFTER: Lips and Face Tracker," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 100-110, San Juan, Puerto Rico, 1997.

[90] B. Schiele and A. Waibel, "Gaze Tracking Based on Face Color," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* Zurich, Switzerland, June 1995.

[91] R. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision, Graphics, and Image Understanding,* vol. 61, no. 1, pp. 38-59, 1995.

[92] D. DeCarlo and D. Metaxas, "The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 231-238, 1996.

[93] T. Jebara, K. Russell, and A. Pentland, "Mixtures of Eigenfeatures for Real-Time Structure from Texture," *Proc. IEEE Int'l Conf. Computer Vision,* Bombay, India, Jan. 1998.

[94] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland, "Visually Guided Graphics," *IEEE Trans. Pattern Analysis and Machine Vision,* vol. 15, no. 6, pp. 602-604, 1993.

[95] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic Modeling for Facial Animation," *Proc. Ann. Conf. Series, SIGGRAPH 1995,* pp. 55-62, 1995.

[96] T. Ishikawa, H. Sera, S. Morishima, and D. Terzopoulos, "Facial Image Reconstruction by Estimated Muscle Parameter," *IEEE Conf. Automatic Face and Gesture Recognition,* pp. 342-347, Nara, Japan, Apr. 1998.

[97] I. Essa and A. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 757-763, July 1997.

[98] H. Li, P. Roivainen, and R. Forchheimer, "3-D Motion Estimation in Model-Based Image Coding," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 6, pp. 545-555, June 1993.

[99] M. Black and Y. Yacoob, "Tracking and Recognizing Rigid and Nonrigid Facial Motion Using Local Parametric Models of Image Motion," *Proc. IEEE Int'l Conf. Computer Vision,* Cambridge, Mass., 1995.

[100] I. Kakadiaris and D. Metaxas, "Model-Based Estimation of 3-D Human Motion Based on Active Multi-Viewpoint Selection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1996.

[101] J. Rehg and T. Kanade, "Visual Tracking of High DOF Articulated Structures: An Application to Human Hand Tracking," *Proc. European Conf. Computer Vision,* vol. II, pp. 35-46, 1996.

[102] E. Di Bernardo, L. Goncalves, and P. Perona, "Monocular Tracking of the Human Arm in 3-D, in Computer Vision for Human-Machine Interaction," R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[103] C. Wren and A. Pentland, "Dynamic Modeling of Human Motion," *Proc. IEEE Conf. Automatic Face and Gesture Recognition,* pp. 22-27, Nara, Japan, Apr. 1998.

[104] C. Bregler, "Tracking People with Twists and Exponential Maps, *IEEE Conf. Computer Vision and Pattern Recogntion,"* 1998.

[105] D. Gavrila and L. Davis, "3-D Model-Based Tracking of Humans in Action: A Multi-View Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 73-80, 1996.

[106] M. Kruger, *Virtual Reality.* Addison-Wesley, 1983.

[107] K. Mase, "Human Reader: A Vision-Based Man-Machine Interface," *Computer Vision Human-Machine Interaction,* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[108] http://www. microsoft. com/billgates/pdc. htm

[109] V. Pavlovic, R. Sharma, and T. Huang, "Gestural Interface to a Visual Computing Enviroment for Molecular Biologists," *Proc. IEEE Int'l Conf. Face and Gesture Recognition,* pp. 30-35, Killington, Vt., Oct. 1996.

[110] M. Black, F. Berard, A. Jepson, W. Newman, E. Saund, G. Socher, and M. Taylor, "The Digital Office: Overview," *Proc. AAAI Spring Symp. Series,* pp. 1-6, Stanford Univ., Mar. 1998.

[111] K. Jo, Y. Kuno, and Y. Shirai, "Manipulative Hand Gesture Recognition Using Task Knowledge for HCI," *Proc. IEEE Conf. Automatic Face and Gesture Recognition,* pp. 468-473, Nara, Japan, Apr. 1998.

[112] S. Stillman, R. Tanawongsuwan, and I. Essa, "System for Tracking and Recognizing Multiple People," *Proc. IEEE Second Int'l Audio- and Video-Based Biometric Person Authentication,* pp. 96-101, Washington, D.C., Mar. 1999.

[113] M. Cohen, "Design Principles for Intelligent Environments," *Proc. AAAI Spring Symp. Series,* pp. 26-43, Stanford Univ., Mar. 1998.

[114] J. Crowley and F. Berard, "Multi-Modal Tracking of Faces for Video Communications," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 640-645, San Juan, Puerto Rico, 1997.

[115] T. Darrell, G. Gordon, J. Woodfill, and M. Harville, "Tracking People with Integrated Stereo, Color, and Face Detection," *AAAI Spring Symp. Series,* pp. 44-50, Stanford Univ., Mar. 1998.

[116] C. Wren, "Understanding Expressive Action," Technical Report 498, Massachusetts Institute of Technology, Media Lab, 1999.

[117] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Shutte, and A. Wilson, "The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment," *Presence,* vol. 8 no. 4, pp. 367-391, 1999.

[118] K. Waters, J. Rehg, M. Loughlin, S. Kang, and D. Terzopoulos, "Visual Sensing for Active Public Spaces," *Computer Vision for Human-Machine Interaction,* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[119] A. Shafer, J. Krumm, B. Brumitt, B. Meyers, M. Czerwinski, and D. Robbins, "The New EasyLiving Project at Microsoft," *Proc. DARPA/NIST Smart Spaces Workshop,* 1998.

[120] Y. Yacoob, L. Davis, M. Black, D. Gavrila, T. Horsrasert, and C. Morimoto, "Looking at People in Action—An Overview," *Computer Vision for Human-Machine Interaction,* R. Cipolla and A. Pentland, eds., Cambridge Univ. Press, 1998.

[121] G. Furnas, T. Landauer, L. Gomes, and S. Dumais, "The Vocabulary Problem in Human-System Communications," *Comm. ACM,* vol. 30, no. 11, pp. 964-972, 1987.

[122] S. Mann, "Smart Clothing: The Wearable Computer and Wear-Cam," *Personal Technologies,* vol. 1, no. 1, 1997.

[123] T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, and A. Pentland, "Visual Augmented Reality Through Wearable Computing," *Presence,* vol. 6, no. 4, pp. 386-398, 1997.

[124] Y. Rosenberg and M. Werman, "Real-Time Object Tracking from a Moving Video Camera: A Software Approach on a PC," *Proc. IEEE Workshop Applications of Computer Vision,* pp. 238-239, Oct. 1998.

Alex Pentland is the academic head of the Massachusetts Institute of Technology (MIT) Media Laboratory, co-founder and co-director of the Center for Future Health at the University of Rochester, and co-founder and vice-chair of the IEEE Computer Society's technical committee on wearable information devices. Newsweek magazine has recently named him one of the 100 Americans most likely to shape the next century and he has won awards from several academic societies, including the AAAI, IEEE, and Ars Electronica. His overall focus is on using digital technology for the worldwide reinvention of health, education, and community. Toward this end, he has done research in wearable computing, human-machine interface, computer graphics, artifical intelligence, machine and human vision, and has published more than 200 scientific articles in these areas.

His research interests include both understanding human behavior (e.g., face, expression, and intention recognition; word learning; and acoustic scene analysis) and wearble computing (e.g., augmenting human intelligence and perception by building sensors, displays, and computers into glasses, belts, shoes, etc.). These are described in the April 1996 and November 1998 issues of *Scientific American*, respectively. Dr. Pentland is a member of the IEEE Computer Society.