AI Magazine Volume 20 Number 2 (1999) (© AAAI)

Computers Seeing People

Irfan A. Essa

■ AI researchers are interested in building intelligent machines that can interact with them as they interact with each other. Science fiction writers have given us these goals in the form of HAL in *2001: A Space Odyssey* and Commander Data in *Star Trek: The Next Generation*. However, at present, our computers are deaf, dumb, and blind, almost unaware of the environment they are in and of the user who interacts with them. In this article, I present the current state of the art in machines that can see people, recognize them, determine their gaze, understand their facial expressions and hand gestures, and interpret their activities. I believe that by building machines with such abilities for perceiving, people will take us one step closer to building HAL and Commander Data.

Building machines that can see has been one of the most exciting and challenging research quests of the last 30 years. Much effort has been expended on "automatic deduction of structure of a possibly dynamic three-dimensional world from two-dimensional images" (Nalwa 1993). There has been considerable progress in the areas of object recognition, image understanding, and scene reconstruction from single and multiple images. This progress, coupled with the improvements in computational power, has prompted a new research focus of making machines that can see people; recognize them; and interpret their gestures, expressions, and actions. In this article, I present methods that give machines the ability to see people, understand their actions, and interact with them. I present the motivating factors behind this work, examples of how such computational methods are developed, and their applications.

The basic reason for providing machines the ability to see people really depends on the task we associate with a machine. An industrial vision system aimed at extracting defects on an assembly line need not know anything about people. Similarly, a computer used for e-mail and text writing need not see and perceive the user's gestures and expressions. However, if our interest is to build intelligent machines that can work with us, support our needs, and be our helpers, then these machines should know more about who they are supporting and helping. If our computers are to do more than support our text-based needs such as writing papers, creating spreadsheets, and communicating by e-mail, perhaps taking on the role of being a personal assistant, then the ability to see a person is essential. Such an ability to perceive people is something that we take for granted in our everyday interactions with each other. This ability to perceive people and interact with them naturally is essential as we move toward building machines like HAL in *2001: A Space Odyssey* and Commander Data in *Star Trek: The Next Generation*.

At present, our model of a machine, or more specifically of a computer, is something that is placed in the corner of the room. It is deaf, dumb, and blind and has no sense of the environment around it or of a person near it. We communicate with this computer using a coded sequence of tappings on a keyboard. Imagine a computer that knows you are near it, knows you are looking at it, and knows who you are and what you are trying to do. Such abilities in a computer are hard to imagine, unless it has an ability to perceive people. Research in speech recognition has made considerable progress toward perception of human speech (see Cole et al. [1995] for a survey). Commercial systems capable of word spotting and recognition of continuous speech are now available. Analysis of the video signal to perceive people has become a challenging and exciting research avenue for the field of computer vision, resulting in significant progress in the recent years.

To make machines that see people, the computer must first determine if someone is near it (where) and count how many people are in its field of view. The next step is to identify who the people are. After the computer has identified the people, it can interpret facial expression, hand gestures, and body language to

determine what the people want or are doing in the scene and why. In the upcoming sections, I present the approaches to determine where, how many, who, what, and why with reference to people in a scene. The answer to each question is not possible independently, and each question depends on the other as dictated by the situation. Before getting into details, I briefly discuss the applications of such a technology.

If our computers are to do more than support our text-based needs such as writing papers, creating spreadsheets, and communicating by e-mail, perhaps taking on the role of being a personal assistant, then the ability to see a person is essential.

Applications

Applications of computer vision methods aimed specifically at seeing people are many and encompass several different areas.

Effective human-computer interaction (HCI): Imagine computers that interact with you as we interact with each other, using speech and gestures. Such computers will know when you are looking at them, will be able to detect where you are pointing, and will interpret your gestures. These types of gestural interface are an integral part of a growing trend toward more human-centered interfaces in HCI research. Specific applications for this technology arise in areas where traditional interfaces such as the keyboard and mouse are not effective. Such techniques will allow us to move toward more noninvasive and unencumbered interfaces that allow for interactive visualization of multidimensional scientific data and user-centered direct interaction with virtual environments.

Smart and interactive environments: Machines that can see will aid us in developing *smart rooms,* rooms that know who is where and what they are doing. Such rooms can help monitor children, senior citizens, or physically challenged individuals and provide assistance and care as needed (Pentland 1996). These types of system and the related interfaces could become a part of our daily activities.

Surveillance and security: A more traditional application of this work is surveillance and security. Face recognition has become quite a useful technology in the security industry, where access is allowed based on facial identity. Systems that automate searches of mug-shot databases to aid in criminal identification are being considered. Recently, work aimed at recognition of human actions promises great help for active surveillance applications.

Entertainment, education, and training: Two areas of recent rapid growth are education and entertainment. Computer vision methods for noninvasive tracking and interpretation of human activities can revolutionize various aspects of these areas too. An intelligent tutor that can see will be far more responsive to the needs of its student if the tutor can judge by the actions and moods of the student whether he/she is confused, frustrated, or confident. Similarly, the development of complex environments for gaming and training will rely on the recognition and interpretation of actions and intentions of the user. A system that understands motions can aid in training for sports and teaching dance.

Video conferencing and model-based coding: Analysis and recognition of facial actions, gestures, and body language, especially with model representations of actions, would be useful for symbolic compression of video data. Vision-based methods for extracting spatiotemporal procedural information of hand gestures, body movements, and facial expressions will aid in the development of model-based video coding methods. With these methods, low–bit-rate videophones and model-based coding systems can be developed. The Moving Picture Experts Group (MPEG) (1999) community is already looking into these issues (see MPEG.org).

Digital libraries and video-image annotations: Automatic content-based annotation of images and video is an important application, especially as the amount of digital content grows at an exponential rate. Because a sizable portion of these data are about people, machines that can recognize people and their activities in images and video will play a significant role in the automatic annotation of these data.

Human augmentation and wearable computing: Systems that can interpret activities of the people in an environment could provide invaluable assistance to hearing-impaired or visually impaired individuals by translating the missing communication modality into a modality that they can directly understand. For example, a seeing computer might describe the body language of a conversational partner to a visually impaired individual through an earphone. The technology could also allow the impaired individual to communicate more effectively, for example, by translating sign language into spoken English (Starner 1995). This form of intelligence driven by perceptual processing and aimed primarily at augmenting users, is becoming an important and challenging research area, especially as computers are taking on newer "roles," for example, wearable computing and affective computing (Wearable 1999; Picard 1998).

In the upcoming sections, I discuss the various aspects of research in computer vision that will play an essential role in the building of machines that can see people.

Is Someone There? Where? (Looking for People)

The first step toward building computers that are aware of people around them is to provide them with the ability to ask, Is someone there? Where? What is their location? Where are they looking? This is achieved by various methods, each varying in detail and function. The most common approaches include subtracting simple backgrounds, looking for specific color features, tracking motions, detecting changes, looking for faces, and tracking heads to determine a pose. I discuss these methods briefly here. First, I address methods for tracking whole bodies from imagery, then I present methods for tracking heads and determining head pose. Whole-body–tracking methods determine where people are, and head-tracking methods extract where people are looking.

People Tracking

The simplest methods for tracking people in a scene are based on image differencing. In these methods, the background image is acquired and stored before the person enters the scene. The person is then segmented in the image by subtracting each new incoming image with the stored background image, which extracts a silhouette of a moving person. A more general method for tracking people using this type of background subtraction requires modeling the scene as a set of distinct classes, including a background class and several classes that cover the person in the foreground.

The PFINDER system uses background and foreground classes to distinguish between the foreground silhouette and the fixed background (Wren et al. 1997). This operation provides the system with a background class while the person is modeled as a connected set of blobs in the foreground, each connected set defining a class. Each blob has spatial (x, y) and color (Y, U, V) properties. In each image of the scene, every pixel must belong to one of the classes. A representation of flesh colors is also encoded to aid in tracking hands and face. These blob features allow tracking of a person's hands and head from low-resolution imagery in real time. To aid in tracking, a low-level description of a model of a person—hands are on the sides, and the head is the highest point of the moving blobs—is also used. The approximate hand positions extracted in this way are used for static gesture recognition. Recent extensions to color-tracking methods include developing Gaussian mixture models of color space to extract flesh tones in the scene.

The basic limitation of the color-based tracking methods is the inherent limitations resulting from the use of color as a metric. Although skin color is a reliable feature for distinguishing between other parts of a person and the hands and face, it has serious problems when users wear skin-colored clothes or short-sleeved shirts and shorts. This limitation is addressed by combining various measurements, as discussed later.

A major advantage of such color-based foreground-background segmentation systems is that they can run in real time on simple desktop computers, allowing for easy development of simple systems for tracking people. Such color-based tracking systems have been demonstrated live during conferences and exhibitions (Darrell et al. 1998; Mase 1993a). However, a significant limitation still exists that current desktop computers barely allow for full-frame video capture (30 frames a second, 640 x 480 pixel resolution) in real time. Most color-tracking systems run at 10 frames a second on a 320 x 240 image and are sufficient for tracking people that don't move too fast.

Another limitation of using a single-camera system running a color-based blob tracker is that it requires a well-calibrated three-dimensional (3D) environment for 3D tracking of the user. This is addressed by running the same blob tracker on two different cameras and extracting positions of the person in 3D using image correspondences, triangulation, and camera-to-camera calibration. A wide-baseline stereo camera system can be used to self-calibrate such a scene, and then stereo matching can be used to track a person in real time. However, it should be obvious that color-tracking methods cannot directly be extended to track multiple people.

Reliable tracking of multiple people is achieved by implementing simple background subtraction techniques in a well-constrained and calibrated closed-world environment. In the KIDSROOM environment (Bobick et al. 1997), a complete domain of the scene being observed is defined, and then silhouettes are tracked over time. With simple metrics of velocity, occlusions are resolved. The domain and the storyboard of an interactive entertainment space are used to determine and control the activities of the participants (users) to aid in tracking.

In addition to the color-based methods, several other methods have been proposed. These methods use a detailed a priori structure of the person being tracked. Similar to the methods discussed earlier, these methods also extract image features—silhouettes, color, and edges—from a scene to aid in tracking people.

The first step toward building computers that are aware of people around them is to provide them with the ability to ask, Is someone there? Where? What is their location? Where are they looking?

Baumberg and Hogg (1994) present methods for using simple models and active contour representations to locate and track people. Bregler and Malik (1998) present a feature-based tracking method coupled with a kinematic model of a walking person to track people. Gavrila and Davis (1996) present a method for tracking people from multiple views. These are more robust tracking techniques compared to the simple color-based tracking methods. In addition, these model-based methods are also more accurate at characterizing the motions. However, these methods are also more computationally complex and require special hardware to achieve real-time performance.

Finding Faces

With the methods described earlier, people are located by simply looking for specific colors or detecting a change in an image. There is no real notion of a person, except when defined a priori to aid in tracking.

A completely different type of method aimed at locating people uses an a priori model of a face and its features to search for a face over the whole image. These methods use features associated with facial shape to determine the number of faces in a scene (Boluja and Kanade 1998a, 1998b; Colmenarez and Huang 1997; Lueng, Burl, and Perona 1995; Moghaddam and Pentland 1995; Turk and Pentland 1991). The techniques that are used in these methods are similar to the ones used in face-recognition methods and are discussed later.

These methods are not yet fast enough for real-time tracking and are presented mostly as a way of extracting faces from static and complex scenes. A real benefit of these systems is that most of these methods are reliable for locating multiple people in a scene. The increase in available computation power will allow for real-time application of these methods. These methods can be combined with the color-tracking or motion-change–detection algorithms to reduce the search space, as discussed later. These systems might serve as a precursor to the face-recognition system that answers the question, Who is in the scene?

Head-Pose Tracking

Determining where a person is and where a person is looking is extremely important for development of systems that are aware of people and are able to recognize the person's face and expressions. Most techniques for expression tracking and face recognition work reliably only for small head motions. This limitation reduces the applicability of these methods, and consequently, head tracking has become an increasingly important research topic.

Head tracking can be achieved by observing a set of features on a face or warping a template of a face to match the transformations of the face as it moves. All the problems inherent in head tracking and pose determination are the same as in determining the orientation of an object for object recognition. Methods that attempt to extract complete 3D structure of the face from visible features to methods that match image templates with affine transformations have been developed. Azarbayejani et al. (1993) present a recursive estimation method for extracting structure and motion of a head by tracking small facial features such as the corners of the eyes or mouth. However, because of its dependence on feature tracking, its applicability is limited to sequences in which the same points were visible over most of the image sequence.

Black and Yacoob (1995) have developed a regularized optical-flow method that uses an 8-parameter 2D affine model of flow that yields good results for pose estimation. However, the use of a planelike 2D model limits accurate tracking to medium-size head motions, and the method can fail for very large head rotations.

Robust head tracking requires a technique that can be characterized as motion regularization or flow regularization (Essa et al. 1996). In this technique, flow between two frames is computed, and the rigid motion of the 3D-head model that best accounts for this computed flow is used as an estimate of head motion. The results of this model-based tracking are shown in figure 1, which shows five frames from a long sequence of a person moving his head.

Combining People Tracking with Face Finding

Robust methods for tracking multiple people using multiple cameras are also being developed. These methods rely on combining methods for color- and silhouette-based tracking with face-detection methods. Stillman, Tanawongsuwan, and Essa (1999) present a robust real-time method for tracking multiple people with multiple cameras. In this method, both static cameras and pan-tilt-zoom (PTZ) cameras are used to extract visual attention in the environment. The PTZ camera system uses face recognition (described in the next section) to register people in the scene and lock on to these individuals. A commercially available face-recognition system (Visionics 1997) that runs in near real time on a PENTIUM PC is used for face tracking and identification. This commercial system uses the video signal from the

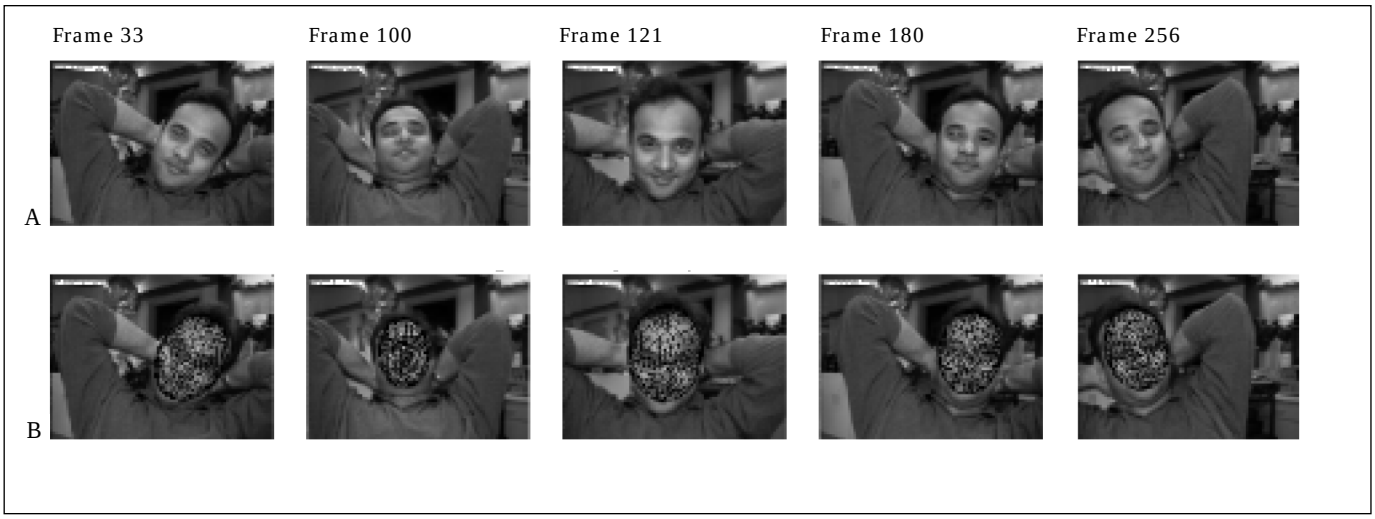


Figure 1. Results of Tracking a Sequence with an Ellipsoidal Model.

A. Original image sequence (300 frames). B. Tracking using three-dimensional ellipsoidal model.

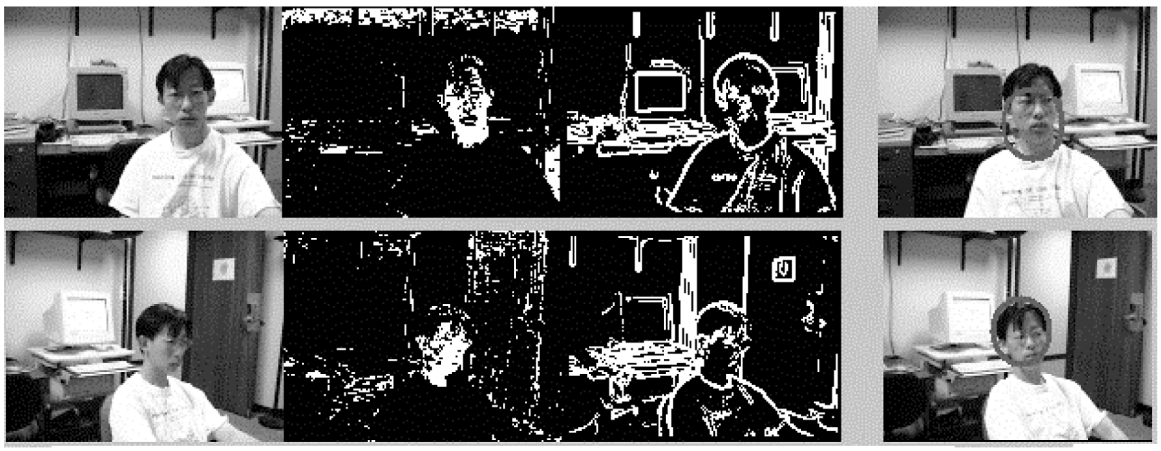


Figure 2. A System Tracking and Following a User's Face.

A combination of color segmentation, movement tracking, and shape information is used for robust tracking of a face.

PTZ cameras to find a face and adjusts the visual foveation process of the PTZ camera.

The static camera system provides a global view of the environment and is used to readjust tracking when the PTZ cameras lose their targets. The system works well even when people occlude one another. The underlying visual processes rely on color segmentation using blob tracking, movement tracking, and shape information to locate target candidates. Color-indexing and motion-tracking modules help register these candidates with the system, allowing for robust tracking. Results of this system are shown in figure 2 for tracking a face using a single camera. The multiple-camera, multiple-people tracking system is described in figures 3 and 4. A distinctive advantage of this type of foveation mechanism is that in addition to a good estimate of the location of the person and his/her face, the system acquires a higher-resolution image of the face that can help with recognition or expression tracking.

Darrell et al. (1998) also present a method that combines face tracking (Rowley, Baluja, and Kanade 1998a) with color tracking to set up a multimodal system for tracking and identifying people.

A B

Figure 3. Results of a Multiple-People, Multiple-Camera Tracking.

A. Two people entering a scene are tracked as they move around. B. The two people occlude each other.

Who Is It? (Recognizing People)

Over the past 30 years, extensive research has been conducted by psychophysicists, psychologists, neuroscientists, and engineers on various aspects of face recognition by humans and machines (see Bruce [1988] and Ellis et al. [1986] for review of work on human perception of faces). The earliest work on machine recognition of faces appeared in the mid-1970s, when typical pattern-classification techniques were used to measure and compare facial-feature attributes for recognition (Kanade 1977). Not much work appeared in this area until the 1990s when the availability of increased computational power, coupled with a commercial demand of face-recognition systems, made the problem computationally viable and commercially exciting.

At present, face recognition is perhaps the most widely studied topic in the vision community. It has the distinct privilege of being the first application of computer vision to be commercialized that is not related to industrial machine vision. At last count, there were 19 commercial ventures attempting to bring face-recognition applications to the public (Face 1999).

The last few years of increased activity have seen progress in locating and segmenting a face in a complex scene; extracting features such as eyes, mouth, and nose; and recognizing occlusions and changes in facial features with orientation, pose, and scale variability. It should be noted that all these problems are standard problems also addressed in the traditional computer vision goal of object recognition and are now being applied to the newer domain of faces. The face-recognition domain, because of its inherent applications, has resulted in significant advances in the design of statistical and neural network–based classifiers. Because of the existence of a large body of literature on a machine-vision method for face recognition, my exposition of this area is brief. Interested readers are encouraged to peruse survey publications by Chellappa, Wilson, and Sirohey (1995) and Samal and Iyengar (1992).

Pattern-Recognition Methods for Face Recognition

As stated earlier, face-recognition methods have resulted in significant developments in various pattern-recognition methods. Recently, a need for a suitable representation for detection and recognition of faces from images has generated renewed interest in Karhunen-Loeve expansion methods (Kirby and Sirovich 1990; Sirovich and Kirby 1987). Karhunen-Loeve expansion methods, also known as principal component analysis (PCA) methods, are widely used in the pattern-recognition area.

A PCA-based method called eigenfaces (Moghaddam and Pentland 1995; Pentland, Moghaddam, and Starner 1994; Turk and Pentland 1991) for face recognition has shown very high recognition accuracy (around 95 percent) using databases of more than 7500 face images of about 3000 people. In this method, faces are aligned with each other and treated as high-dimensional pixel vectors from which eigenvectors (called *eigenfaces*) and eigenvalues are computed. These eigenvectors represent the principal components; therefore, the eigenvalue decomposition method allows for representing the probe face by a small number (some-

Figure 4. Views from Two Static Cameras (Left and Right) Showing the Result from the Person-Detection System.

Also shown are views from the two pan-tilt-zoom (PTZ) cameras placed in the front of the room. A triangulation process is used to decide the scale factor (that is, the distance from a person to the PTZ camera).

times 100) of expansion coefficients, which are then used in recognition. The alignment of all the faces is done automatically by using a similar representation for facial features (eyes, nose, and mouth). Several extensions to these methods have recently been proposed (Etemad and Chellappa 1994; Swets and Weng 1996).

Another popular method relies on collapsing the variances in facial images to extract face descriptors called *image graphs*. In these graphs, facial features are described as a set of wavelet components. Image graphs are extracted by generating a *bunch graph*, which is constructed from a small set of sample image graphs. Comparison of this image graph between images yields recognition of facial images (Kruger, Potzsch, and von der Malsburg 1997; Wiskot et al. 1997). This work extends the work of Manjunath, Chellappa, and von der Malsburg (1992) that uses Gabor wavelet decomposition and that of Landes et al. (1993) that uses dynamic link architecture (DLA). Impressive results for recognition of faces from different viewpoints are reported.

In addition to classical pattern-recognition methods, much work exists on applications of neural networks for face recognition. Rowley, Baluja, and Kanade (1998a, 1998b) present good results for face detection using retinally connected neural nets that examine small windows of an image and decide whether each window contains a face. They use multiple neural nets and have shown reliable results with large variations in pose. Brunelli and Poggio (1993) present a different method using a HYPERBF network for recognition of a face.

Because of the large body of work on face recognition in recent years, it is almost impossible to cover all the significant developments. However, it is important to observe that each system claims good results, and the authors freely discuss the strengths and weaknesses of each method. Until recently, there was no definitive way of comparing these results, which led to the Face Recognition Technology Program (FERET) evaluation sponsored by the United States Department of Defense. Phillips et al. (1998, 1997) present the methodology and the results of these tests. The FERET Program provides a methodology for reliable testing of different face-recognition systems over a large database (14,126 images of 1,199 people) collected independently. These tests are very successful in evaluating the state of the art in face-recognition methodologies and measure algorithmic performance over large databases. These tests rated the systems from the Massachusetts Institute of Technology (Moghaddam and Pentland 1995; Pentland, Moghaddam, and Starner (1994), the University of Maryland (Etemad and Chellappa 1996; Manjunath, Chellappa, and von der Malsburg 1992), the University of Southern California (Kruger, Potsch, and von der Malsburg 1997; Wiskot et al. 1997), and Michigan State University (Swets and Weng 1996) as very proficient in recognizing faces.

Does that mean that face recognition is a solved problem? The evidence supports this to be true for face recognition in limited domains and applications with full frontal faces.

Under constrained environments with full-frontal faces, there is every reason to expect these face-recognition systems to perform reliably. However, much research is still needed to resolve face recognition in unconstrained environments with variations in lighting, orientations, and changes in facial features.

What Do They Want or What Are They Doing? (Gesture, Expression, and Activity Recognition)

Now, I present the methods for asking questions of what is happening in an environment. I start with a discussion about recognizing facial expressions, then explore gesture recognition and interpretation of human activity.

Facial Expression Recognition

At present, face recognition is perhaps the most widely studied topic in the vision community. It has the distinct privilege of being the first application of computer vision to be commercialized that is not related to industrial machine vision.

The psychology community has a large body of work on face perception and facial analysis. Perhaps the most important work in this area is the effort led by Ekman, and Friesen (1978), who produced a system for describing all visually distinguishable facial movements called the facial action coding system (FACS). In this system, each expression can be represented in terms of action units. It is believed that automatic recognition of facial expressions from images can be achieved by categorizing a set of predetermined facial motions, such as with FACS, rather than determining the motion of each facial point independently.

Yacoob and Davis (1994), Black and Yacoob (1995), and Mase (1993b) use the FACS representation for recognition of facial expressions. Yacoob and Davis extend the work of Mase by detecting motion in six predefined and hand-initialized rectangular regions on a face and then use simplifications of the FACS rules for the six universal expressions for recognition. The motion in these rectangular regions from the last several frames is correlated to the FACS rules for recognition. Black and Yacoob extend this method further by using local parameterized models of image motion to handle large-scale head motions. These methods show about 89-percent accuracy in correctly recognizing expressions over their database of 105 expressions. They have also shown remarkable success at recognizing expressions from real video of people in television talk shows. These results are impressive considering the complexity of the FACS model and the difficulty in measuring facial motion within small-windowed regions of the face.

It has been argued that one of the main difficulties these researchers have encountered is the complexity of describing human facial movement using FACS. These limitations of FACS as a representation of facial motion for automatic recognition have recently generated a lot of discussion. National Science Foundation (NSF) workshops and the resulting reports on facial expressions discuss this issue in detail (Pelachaud, Badler, and Viaud 1994; Ekman et al. 1993).

Essa and Pentland (1995, 1994) and Essa, Darrell, and Pentland (1994) undertake detailed experiments for measuring facial motion and report that it is important to move away from a static, dissect-every-change analysis of expressions. This extension toward a whole-face analysis of facial dynamics in motion sequences is even more significant for machine perception of facial motion. They have analyzed video data of facial expressions and then probabilistically characterized the facial muscle activation associated with each expression. This characterization is achieved using a detailed, physically based dynamic model of the skin and muscles coupled with optimal estimates of optical flow in a feedback-controlled framework. A second, simpler representation that encodes the motion and velocity in the image plane is also extracted. There are 2D spatiotemporal templates to represent facial expressions.

This detailed analysis of video data yields two representations of facial motion that are then used to recognize facial expressions in two different ways. These extracted representations are graphically shown in figure 5. Both of these methods for recognition of facial expressions result in 98-percent accuracy over 52 image sequences. However, these results are preliminary, and comparison with other techniques is not possible without using all the proposed methods for facial-expression recognition on the same test set. A FERET type of initiative would be beneficial to this type of research.

One of the major problems with these facial-expression techniques is that they do not run in real time or even at interactive rates. At present, the method that uses a dynamic physics-based model of the face is computation intensive. On a Silicon Graphics INDY R5000 180-megahertz machine, each frame takes about 15 seconds. The method that uses the 2D spatiotemporal templates is much more efficient and runs about 5 frames a second (fps) (that is, 1 second of 160- x 120-resolution video at 15 fps would take 3 seconds after digitization). Using specialized hardware and multiple-processor PENTIUMs could aid in such computations.

Essa, Darrell, and Pentland (1994) and Darrell, Essa, and Pentland (1996) present a method for facial tracking and interactive animation of faces that runs in real time. The basic idea for this method is to do a fine-grained analysis of a subject's expression and then store the spatiotemporal representation

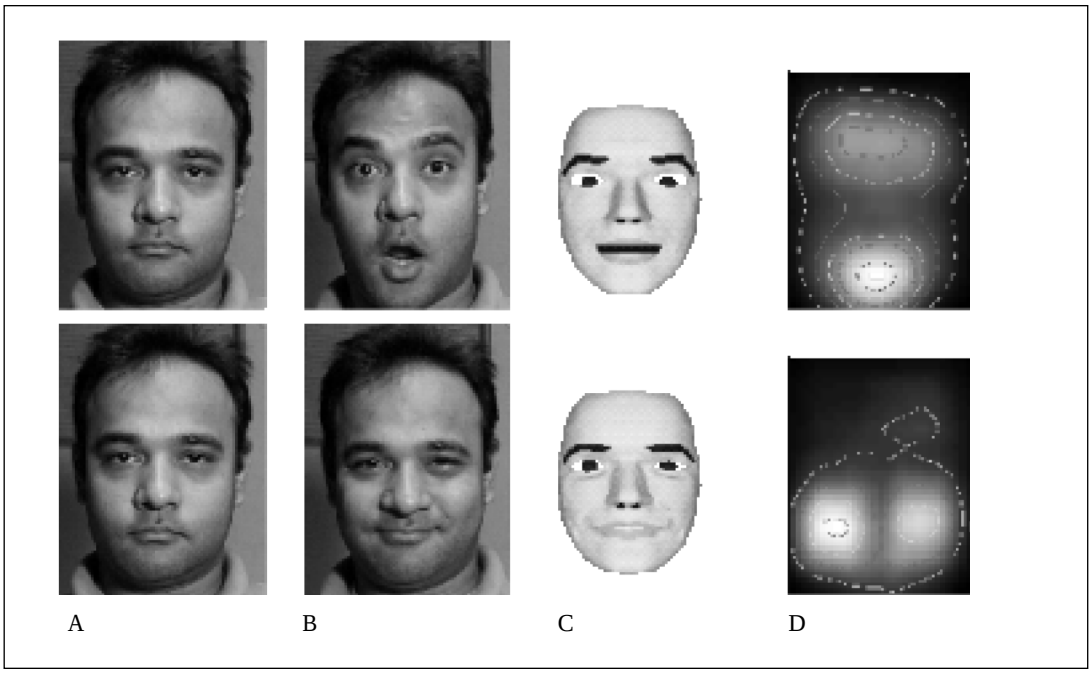


Figure 5. Determining Expressions from Video Sequences.

A. Neutral: The surprise expression showing in the top row, and the smile-happiness expression showing in the bottom row. B. Expression: The model used for analysis synthesis. C. Model. D. The motion energy: Peak muscle actuation and motion energy are used for recognition of expressions (Essa and Pentland 1997).

of this expression on a generic model of a face. Then simple visual measurements can be used to establish the relationship between an image and the dynamic motion parameters of the model. These simple visual measurements could be appearance and view based, feature based, blob based, or even motion based. These measurements are coupled with the parameters of the physics-based model using an interpolation process, resulting in a real-time, passive (that is, the observations drive the model) facial tracking and animation system (figure 6). In addition to tracking expressions using this method, hidden Markov models (HMMs) could be used for recognition of expressions based on a similar set of visual measurements.

It is important to note here that the previous methods are aimed at recognition of facial expressions. Because there is a known relationship between facial expression and human emotions (see Ellis et al. [1986], Bruce [1988], and Ekman and Friesen [1969]), it is foreseeable that such techniques can be used to recognize human emotions. Although the possibilities of developing such systems are both exciting and challenging (Picard 1998) and raise many intriguing social implications, not much work to date has been attempted to build and evaluate such a system. Building machines that can recognize emotions and read lips is an actual goal of the work on the recognition of facial expressions.

Gesture Recognition

There are many facets to modeling, tracking, and recognizing human gesture and body motion. For example, gestures can be made by hands, faces, or the entire body; have strong spatial and temporal characteristics; can be person or culture specific; can be tied to a linguistic basis or spoken conversations; or can be meaningful in their own right. For this reason, research in several domains (vision, AI, linguistics, biomechanics, and robotics) is relevant for automatic understanding of gestures.

Many researchers in the vision community have attempted automatic gesture recognition and body tracking from video (Darrell, Essa, and Pentland 1996; Baumberg and Hogg 1994; Kakidiaris, Mataxas, and Bajcsy 1994; Rehg and Kanade 1994). In these efforts, pattern-recognition methods are applied to extract spatiotemporal codings from image streams for recognition. Learning algorithms have also been used for interpretation of gestures (Starner, Weaver, and Pentland 1998; Yamato, Ohya, and Ishii 1994). This area of research has been furthered by successful attempts at using appearance-based or view-based methods for tracking and recognizing human motion (Black and Jepson 1996; Moghaddam and Pentland 1995).

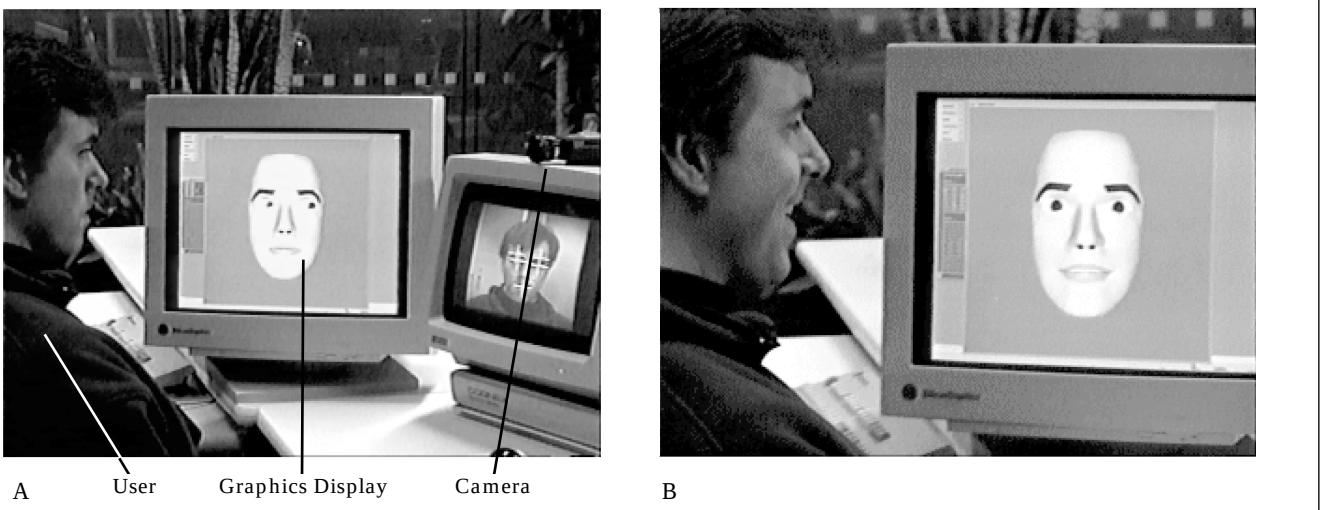


Figure 6. Real-Time Tracking of Facial Movements.

A. Complete system tracking eyes, mouth, eyebrows. B. Mimicking a smile expression.

The importance of time in the analysis and recognition of hand and body movements has led to the use of HMMs for recognition after training on views of the model. For example, Yamato, Ohya, and Ishii (1994) studied the recognition of tennis strokes by training on time-sequential images of six different tennis shots. Starner, Weaver, and Pentland (1998) use HMMs for recognition of American Sign Language. Bobick et al. (1997) and Bobick and Wilson (1995) have shown a unique way of representing gesture that captures both the repeatability and variability of gestures in a training set of example trajectories of gesture states.

Gesture understanding requires interpretation of the spatiotemporal patterns extracted from video with the constraints imposed by the dynamic representation of human action and the linguistic context, if any, of such an action. To achieve such an understanding of human gestures, we need to develop a theory of human action that has an inherent computational value. Essa and Pentland (1995) present a similar idea that relies on a computational value for interpretation of facial expressions. In this approach, a reduced dimensional representation of facial action is developed by a causal reconstruction of how the scene was produced. This representation is achieved by coding facial movements from video in terms of muscle contractions and using an analysis-synthesis framework. A similar attempt at a preliminary extension of this method (framework) for whole-body actions by using a kinematic model of a human figure is presented by Brand and Essa (1995). Lakoff and Johnson's (1980) theory on metaphors for actions is used for empirically defining high-level human actions from low-level kinematic motions.

Tracking three-dimensional human movements from video is far from a trivial problem because tracking generally is an underconstrained problem, the data are noisy, and the measurements include several levels of nonlinearity. Adding a layer of constraints imposed by the dynamic representation of human action and the linguistic context of the action should help with analysis and interpretation.

If the interest is in recognizing and representing higher-level human actions, we can gain insight from research on how humans express themselves and how they move. Unlike the machine-vision community, the linguistics community has been studying the communicative aspects of gestures for many years (Kendon 1974; Ekman and Friesen 1969; Efron 1941). Some recent work is aimed at understanding gestures in the context of communication, especially speech (McNeill 1992; Krauss, Morrel-Samules, and Colasante 1991; Cassell and McNeill 1990). We believe that this work provides us with at least a preliminary understanding of communication through gestures and should provide rules to help with the interpretation of gestures.

For more detailed analysis of human movement, we can rely on the biomechanics literature that provides motion-capture data, force-

plate data, and muscle-activation records. These data tell us how and why humans move in the ways they do. These data can be used to tune control algorithms for human motion and provide additional constraints on the candidate descriptions for a motion sequence.

In computer animation, researchers have explored (Hodgins 1998) the use of dynamic simulation as a technique for generating human motion for computer animation and virtual environments. These dynamic motion generators for human action can be extended to provide us with both a higher-level representation (behavior or activity level) and a lower-level description in space and time (joint angles, positions, and so on) for additional behaviors that are appropriate for any application domain. Additional behaviors such as sitting, walking, pointing, dancing, and gesturing will force us to address stylistic issues. Studying the use of this type of generated motion with an appearance- and motion-based extraction of events from video will yield interesting results.

Activity Recognition

As stated previously, computer vision is a critical technology for creating systems that can interact naturally and intelligently with people. In addition to finding, tracking, and recognizing people, we can use computer vision techniques to recognize human activities in an environment (Seitz and Dyer 1997; Bobick 1996; Polana and Nelson 1993). Such recognition of human activities requires the study of the dynamic relationship between human motion and objects in the scene. Additionally, to address the issue of recognition of human actions and activities, it seems essential to develop an adaptive approach that uses context as a means of deciding the most appropriate representation that will be used for recognition.

It seems apparent that understanding the dynamics of human motion is fundamental to solving action-recognition problems (see Cedras and Shah [1995] for a review). A common thread in much of the recent work in action recognition has been the use of HMMs as a means of modeling complex actions. Lately, there have been several contributions in the literature that offer new frameworks for activity recognition. Specifically, Bregler (1997) evaluates motion at graduated levels of abstraction by using a four-level decomposition framework that learns and recognizes human dynamics in video sequences. Although Bregler's method focuses on complex human motions, such as walking, Oliver, Pentland, and Berard (1997) present a system designed to assess interactions between people using Bayesian approaches. Bobick (1996) also presents several approaches to the machine perception of motion and discusses the role and levels of knowledge in each. The framework proposed by Buxton and Gong (1995) uses Bayesian networks for surveillance activities in well-defined and constrained scenarios.

Context management plays a critical role in this process by supplying, maintaining, and discovering information about the relationships between people and objects. Objects provide clues about which human motions to anticipate, making them powerful tools for discriminating between actions and activities. Building a formal context model for people and their surroundings provides an architecture where acquired visual data can be warehoused, analyzed, and shared effectively.

To address this issue, Moore, Essa, and Hayes (1999) are developing an object-oriented approach called OBJECTSPACES to encapsulate context into scene objects. Instead of making static assumptions about the contents of an image sequence, they attach regions in an image of a scene to virtual objects. A scene object is derived as an instance of a class type. For example, if the scene is an office environment, classes would include desk, bookcase, and keyboard. All scene objects are provided with a priori information about the image regions they represent. By monitoring these regions, objects can develop an awareness of their features and can detect when their state changes. For example, if a person moves a book resting on a table by a few inches, the book object can determine that it has been moved and attempt to recalculate its new position. Additionally, each object understands complex actions that are indigenous to its class. For example, the book object stores a profile of two motion gestures—(1) page forward and (2) page backward—that it can identify by observing how humans interact with it. By evaluating these two actions over time, the book can decide if someone is quickly browsing through its pages or carefully studying every word. Tracking and motion analysis, which takes place in the extraction layer, is shared among objects representing scene articles and people. The scene's objects report their observations to a scene-level object, or *scene layer*, that catalogs all the activities taking place. This layer searches for correlations between object interactions to classify particular activities or identify certain human behaviors.

To test these representations, experiments in natural environments, where people interact with their surroundings, are recorded. The first

The psychology community has a large body of work on face perception and facial analysis. Perhaps the most important work in this area is the effort led by Ekman, and Friesen (1978), who produced a system for describing all visually distinguishable facial movements called the facial action coding system (FACS).

Figure 7. Recognition of Activities around a Workstation.

Using the ObjectSpace representation, objects are marked and their context established. Hands are tracked using color features, and activities are recognized.

experiment is based in a real office environment equipped with typical objects and appliances. Objects in the scene and the action associations are predetermined. Testing on a 5-minute video sequence of a user interacting with different objects in the office scene, the system detected 92 percent of the events correctly (figure 7).

This type of research effort, which is aimed at the recognition of human activity, has many practical applications where passive, nonintrusive action recognition is desired, such as video surveillance and activity annotation. Moreover, work conducted in this area advances computer awareness, which is an essential step toward the building of intelligent machines that can perceive and communicate with us naturally.

Conclusions

In recent years, there has been significant progress in the building of machines that are aware of the users who interact with them. The increase in computational power, combined with the multimedia capabilities of computers, has had a strong impact on the growth of this research area. Development of computer vision systems that are more than simple prototypes, and have applications beyond industrial vision, has been a boon for computer vision. In fact, it has resulted in a significant growth in computer vision research in recent years, mostly supported by industrial funding in addition to the more traditional government funding. It is important to also note that current multimedia computers are also making it very easy to develop vision-based systems for interaction. Applications of this type of computer vision research are many and far reaching.

On a technical level, this domain of computer vision research has revived the concepts of pattern recognition for interpretation of a scene. Face-recognition methods are a perfect example of this revival and are aimed at the static interpretation of an image. In addition, some of the recent work in gesture and action recognition requires a study of dynamic signal and symbol interpretation. Research in several domains (vision, AI, linguistics, biomechanics, and robotics) is essential and relevant if we are interested in building machines that can see us.

These are indeed exciting and challenging problems and exciting and challenging times for research in computer vision. The day is not far away when our desktop computers will be able to see when we are looking at them. The next step toward building HAL and Commander Data is to make these systems robust to varying conditions and responsive to us in real time.

Acknowledgments

Special thanks to the students who have been involved in these projects at Georgia Tech, specifically the students in the Computational Perception Laboratory: G. Brostow, D. Moore, W. Rungsarityotin, A. Schoedl, D. Steedly, S. Stillman, A. Stoytchev, K. Sukel, and R. Tanawongsuwan. Also, thanks to Sandy Pentland (MIT Media Lab) under whose guidance some of the earlier work was undertaken, and thanks to our collaborators M. Brand (Mitsubishi Electric Research Labs), S. Basu (MIT), T. Darrell (Interval), and A. Ram (Georgia Tech).

References

Azarbayejani, A.; Horowitz, B.; and Pentland, A. 1993. Recursive Estimation of Structure and Motion Using the Relative Orientation Constraint. In Proceedings of the Computer Vision and Pattern Recognition Conference, 15–17 June, New York.

Azarbayejani, A.; Starner, T.; Horowitz, B.; and Pentland, A. P. 1993. Visually Controlled Graphics. *IEEE Transactions on Pattern Analysis* 15(6): 602–605.

Baumberg, A., and Hogg, D. 1994. An Efficient Method for Contour Tracking Using Active Shape Models, TR-94.11, School of Computer Studies, University of Leeds.

Black, M. J., and Jepson, A. D. 1996. EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. *International Journal of Computer Vision* 26(1): 3–84.

Black, M. J., and Yacoob, Y. 1995. Tracking and Recognizing Rigid and Non-Rigid Facial Motions Using

Local Parametric Model of Image Motion. In Proceedings of the International Conference on Computer Vision, 374–381. Washington, D.C.: IEEE Computer Society.

Bobick, A. F. 1996. Computers Seeing Action. Technical Report, 394, Media Laboratory, Perceptual Computing Section, Massachusetts Institute of Technology.

Bobick, A. F., and Wilson, A. D. 1995. A State-Based Technique for the Summarization and Recognition of Gesture. In Proceedings of the International Conference on Computer Vision. Washington, D.C.: IEEE Computer Society.

Bobick, A.; Intille, S.; Davis, J.; Baird, F.; Pinhanez, C.; Campbell, L.; Ivanov, Y.; Schutte, A.; and Wilson, A. 1997. The KIDSROOM: A Perceptually Based Interactive and Immersive Story Environment. Technical Report, 398, Media Laboratory, Perceptual Computing Section, Massachusetts Institute of Technology.

Brand, M., and Essa, I. 1995. Causal Analysis for Visual Gesture Understanding. Paper presented at the AAAI Fall Symposium on Computational Models for Integrating Language and Vision, 10–12 November, Cambridge, Massachusetts.

Bregler, C. 1997. Learning and Recognizing Human Dynamics in Video Sequences. In Proceedings of the Computer Vision and Pattern Recognition Conference, 568–574. Washington, D.C.: IEEE Computer Society.

Bregler, C., and Malik, J. 1998. Tracking People with Twists and Exponential Maps. In Proceedings of Computer Vision and Pattern Recognition, 8–15. Washington, D.C.: IEEE Computer Society.

Bruce, V. 1988. *Recognizing Faces*. Mahwah, N.J.: Lawrence Erlbaum.

Brunelli, R., and Poggio, T. 1993. Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(10): 1042–1052.

Buxton, H., and Gong, S. 1995. Advanced Visual Surveillance Using Bayesian Networks. In Proceedings of the IEEE Workshop on Context-Based Vision. Washington, D.C.: IEEE Computer Society.

Cassell, J., and McNeill, D. 1990. Gesture and Ground. In Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society, eds. K. Hall, J.-P. Keonig, M. Meachman, S. Reinman, and L. Sutton, 57–68. Berkeley, Calif.: Berkeley Linguistics Society.

Cedras, C., and Shah, M. 1995. Motion-Based Recognition: A Survey. *Image and Vision Computing* 13(2): 129-155.

Chellappa, R.; Wilson, C. L.; and Sirohey, S. 1995. Human and Machine Recognition of Faces: A Survey. *Proceedings of IEEE* 83(5): 705–740.

Colmenarez, A. J., and Huang, T. S. 1997. Face Detection with Information-Based Maximum Discrimination. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Conference 1997, 782–787. Washington, D.C.: IEEE Computer Society.

Darrell, T.; Essa, I.; and Pentland, A. 1996. Task-Specific Gesture Analysis in Real Time Using Interpolated Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(12): 1236–1242.

Darrell, T.; Gordon, G.; Harville, M.; and Woodfill, J. 1998. Multi-Modal Person Detection and Identification for Interactive Systems. In Proceedings of Computer Vision and Pattern-Recognition Conference. Washington, D.C.: IEEE Computer Society.

Efron, D. 1941. *Gesture and Environment*. New York: King's Crown.

Ekman, P., and Friesen, W. 1969. The Repertoire of Nonverbal Behavioral Categories—Origins, Usage, and Coding. *Semiotica* 1:49–98.

Ekman, P., and Friesen, W. V. 1978. *Facial Action Coding System*. Palo Alto, Calif.: Consulting Psychologists Press.

Ekman, P.; T. Huang, T.; Sejnowski, T.; and Hager, J., eds. 1993. Final Report to NSF of the Planning Workshop on Facial Expression Understanding. Technical report, National Science Foundation, Human Interaction Lab, University of California at San Francisco.

Ellis, H. D.; Jeeves, M. A.; Newcombe, F.; and Young, A., eds. 1986. *Aspects of Face Processing*. Zoetermeer, The Netherlands: Martinus Nijhoff.

Essa, I., and Pentland, A. 1995. Facial Expression Recognition Using a Dynamic Model and Motion Energy. In Proceedings of the International Conference on Computer Vision, 360–367. Washington, D.C.: IEEE Computer Society.

Essa, I., and Pentland, A. 1994. A Vision System for Observing and Extracting Facial Action Parameters. In Proceedings of the Computer Vision and Pattern Recognition Conference, 76–83. Washington, D.C.: IEEE Computer Society.

Essa, I.; Darrell, T.; and Pentland, A. 1994. Tracking Facial Motion. In Proceedings of the Workshop on Motion of Nonrigid and Articulated Objects, 36–42. Washington, D.C.: IEEE Computer Society.

Essa, I.; Basu, S.; Darrell, T.; and Pentland, A. 1996. Modeling, Tracking, and Interactive Animation of Faces and Heads Using Input from Video. In Proceedings of Computer Animation Conference 1996, 68–79. Washington, D.C.: IEEE Computer Society.

Etemad, K., and Chellappa, R. 1994. Face Recognition Using Discriminant Eigenvectors. Paper presented at the International Conference on Acoustics, Speech, and Signal Processing, 19–22 April, Adelaide, Australia.

Face Recognition. 1999. Face Recognition Web Page. Available at www.cs.rug.nl/~peterkr/FACE/face.html.

Gavrila, D. M., and Davis, L. S. 1996. 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 73–80. Washington, D.C.: IEEE Computer Society.

Hodgins, J. 1998. Animating Human Motion. *Scientific American* 276(3).

Kakadiaris, I.; Metaxas, D.; and Bajcsy, R. 1994. Active Part-Decomposition, Shape, and Motion Estimation of Articulated Objects: A Physics-Based Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 980–984. Washington, D.C.: IEEE Computer Society.

Kanade, T. 1977. *Computer Recognition of Human Faces*. Cambridge, Mass.: Birkhauser Verlag.

Kendon, A. 1974. Movement Coordination in Social Interaction: Some Examples Described. In *Nonverbal Communication*. New York: Oxford University Press.

Kirby, M., and Sirovich, L. 1990. Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *Pattern Analysis and Machine Intelligence* 12(1): 103–108.

Krauss, R.; Morrel-Samules, P.; and Colasante, C. 1991. Do Conversational Hand Gestures Communicate? *Journal of Personality and Social Psychology* 61(5): 743–754.

Kruger, N.; Potzsch, M.; and von der Malsburg, C. 1997. Determination of Face Position and Pose with a Learned Representation Based on Labeled Graphs. *Image and Vision Computing* 15(8): 665–673.

Kruizinga, P. 1999. Face-Recognition Web Page. Available at www.cs.rug.nl/peterkr/FACE/face.html/.

Landes, B.; Vorbruggen, C.C.; Buhmann, J.; Lange, J.; von der Malsburg, C.; Wurtz, R. P.; and Konen, W. 1993. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers* 42(3): 300–311.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: Chicago University Press.

Leung, T.; Burl, M.; and Perona, P. 1995. Finding Faces in Cluttered Scenes Using Labelled Random Graph Matching. In Proceedings of the International Conference on Computer Vision, 637–644. Washington, D.C.: IEEE Computer Society.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Manjunath, B. S.; Chellappa, R.; and von der Malsburg, C. 1992. A Feature-Based Approach to Face Recognition. In Proceedings of Computer Vision and Pattern Recognition, 373–378. Washington, D.C.: IEEE Computer Society.

Mase, P. 1993a ALIVE: An Artificial Life Interactive Video Environment. In ACM SIGGRAPH Visual Proceedings, 189. New York: Association of Computing Machinery.

Mase, K. 1993b. Recognition of Facial Expressions for Optical Flow. *IEICE Transactions* (Special Issue on Computer Vision and Its Applications) E74(10).

Moghaddam, B., and Pentland, A. 1995. Probabilistic Visual Learning for Object Detection. In Proceedings of the International Conference on Computer Vision. Washington, D.C.: IEEE Computer Society.

Moore, D.; Essa, I.; and Hayes, M. 1999. Context Management for Human Activity Recognition. Paper presented at the Audio-Vision–Based Person Authentication Conference, 22–23 March, Washington, D.C.

MPEG. 1999. Moving Picture Experts Group Web Page. Available at www.mpeg.org/.

Nalwa, V. 1993. *A Guided Tour of Computer Vision*. Reading, Mass.: Addison Wesley.

Oliver, N.; Pentland, A. P.; and Berard, F. 1997. LAFTER: Lips and Face Real-Time Tracker. In Computer Vision and Pattern Recognition, 123–129. Washington, D.C.: IEEE Computer Society.

Pelachaud, C.; Badler, N.; and Viaud, M. 1994. Final Report to NSF of the Standards for Facial Animation Workshop. Technical Report, National Science Foundation, Washington, D.C.

Pentland, A. 1996. Smart Rooms. *Scientific American* 274(4): 68–76.

Pentland, A.; Moghaddam, B.; and Starner, T. 1994. View-Based and Modular Eigenspaces for Face Recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, 84–91. Washington, D.C.: IEEE Computer Society.

Phillips, P. J.; Moon, H. J.; Rauss, P.; and Rizvi, S. A. 1997. The FERET Evaluation Methodology for Face-Recognition Algorithms. In Proceedings of Computer Vision and Pattern Recognition, 137–143. Washington, D.C.: IEEE Computer Society.

Phillips, P. J.; Wechsler, H.; Huang, J.; and Rauss, P. J. 1998. The FERET Database and Evaluation Procedure for Face-Recognition Algorithms. *Image and Vision Computing* 16(5): 295–306.

Picard, R. 1998. *Affective Computing*. Cambridge, Mass.: MIT Press.

Polana, R., and Nelson, R. 1993. Detecting Activities. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2–7. Washington, D.C.: IEEE Computer Society.

Rehg, J. M., and Kanade, T. 1994. Visual Tracking of High DOF Articulated Structures: An Application to Human Hand Tracking. Paper presented at the Third European Conference on Computer Vision, 2–6 May, Stockholm, Sweden.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1998a. Neural Network–Based Face Detection. *Pattern Analysis and Machine Intelligence* 20(1): 23–38.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1998b. Rotation Invariant Neural Network–Based Face Detection. In Proceedings of Computer Vision and Pattern Recognition. Washington, D.C.: IEEE Computer Society.

Samal, A., and Iyengar, P. A. 1992. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey. *Pattern Recognition* 25(1): 65–77.

Seitz, S., and Dyer, C. 1997. View-Invariant Analysis of Cyclic Motion. *International Journal of Computer Vision* 25(3).

Sirovich, L., and Kirby, M. 1987. Low-Dimensional Procedure for the Characterization of Human Faces. *Journal of the Optical Society of America* 4(3): 519–524.

Starner, T. 1995. The MIT Wearable Computing Web Page. Available at wearables.www.media.mit.edu/projects/wearables/.

Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Videos. *IEEE Transactions on Pattern Analysis and Machine Vision* 20(12): 1371–1375.

Stillman, S.; Tanawongsuwan, R.; and Essa, I. 1999. A System for Tracking and Recognizing Multiple People with Multiple Cameras. Paper presented at the Audio- and Vision-Based Person Authentication Conference, 22–23 March, Washington, D.C.

Swets, D. L., and Weng, J. J. 1996. Using Discriminant Eigenfeatures for Image Retrieval. *Pattern Analysis and Machine Intelligence* 18(8): 831–836.

Turk, M., and Pentland, A. 1991. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3(1): 71–86.

Visionics. 1997. FACEIT Developer Kit Version 2.0. Visionics Corporation. Available at www.visionics.com.

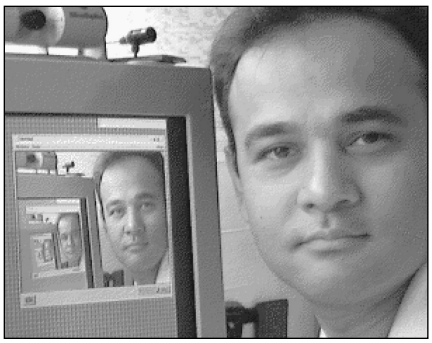
Wiskott, L.; Fellous, J. M.; Kruger, N.; and von der Malsburg, C. 1997. Face Recognition by Elastic Bunch Graph Matching. *Pattern Analysis and Machine Intelligence* 19(7): 775–779.

Wren, C.; Azarbayejani, A.; Darrell, T.; and Pentland, A. 1997. PFINDER: Real-Time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7): 780–785.

Yacoob, Y., and Davis, L. 1994. Computing Spatio-Temporal Representations of Human Faces. In Proceedings of the Computer Vision and Pattern Recognition Conference, 70–75. Washington, D.C.: IEEE Computer Society.

Yamato, J.; Ohya, J.; and Ishii, K. 1994. Recognizing Human Action in Time-Sequential Images Using a Hidden Markov Model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 379–385. Washington, D.C.: IEEE Computer Society.

Wearable Computing. 1999. Wearable Computing Web Page. Available at wearables.www.media.mit.edu/projects/wearables.



Irfan A. Essa is an assistant professor and Imlay fellow in the College of Computing and adjunct assistant professor in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. At Georgia Tech, he is affiliated with the Future Computing Environments effort; the Graphics, Visualization, and Usability Center; and the Intelligent Systems Group in the College of Computing. He has founded the Computational Perception Laboratory that aims to explore and develop the next generation of intelligent machines, interfaces, and environments that can perceive, recognize, anticipate, and interact with humans.

Prior to joining Georgia Tech, Essa was a research scientist in the Perceptual Computing Section of the Media Lab at the Massachusetts Institute of Technology (MIT). He received his Ph.D. from MIT in September 1994. His dissertation dealt with visual analysis and interpretation of facial expressions. He earned his M.S. from MIT (Media Lab and IESL) in 1990 and his B.S. from the Illinois Institute of Technology. His web page is located at www.cc.gatech.edu/~irfan, and his e-mail address is irfan@computer.org.