

Face and Gesture Recognition: Overview

John Daugman, *Associate Editor, PAMI*

FACE and gesture recognition are effortless aspects of interaction among humans, but human-computer interaction remains based upon signals and behaviors which are not natural for us. Although keyboard and mouse are undeniable improvements over tabular switch and punch card, the development of more natural and intuitive interfaces that do not require humans to acquire specialized and esoteric training remains an elusive goal. To facilitate a more fluid interface, we would prefer to enable machines to acquire human-like skills, such as recognizing faces, gestures, and speech, rather than continuing to demand that humans acquire machine-like skills.

No doubt the most important advancements towards these goals will come from natural language interfaces that provide continuous speech processing, speaker-independent recognition, and natural-sounding speech synthesis. But an important adjunct will also come from robust ways to process, and indeed to emulate, non-verbal human expression such as manual or facial gesticulation. One benefit is that we naturally use our faces as organs of expression, reaction, and query, and also to modulate visually or to add nuance to speech. Facial expression and head movements are often used even in lieu of speech when we express assent, dissent, uncertainty, irony, and other reactive states in face-to-face conversation with each other.

International standards organizations that define electronic communication and compression protocols already specify provision in the near future for efficient coding and synthesis of human gesture and facial expression in the transmission of video data. For example, the international MPEG-4 standard is due to include provision for facial animation by January 1999, and body animation beyond 1999. These capabilities are important for low-bandwidth transmission of video data involving remotely interacting persons, as in a teleconference or other multimedia context. A raw video stream (bandwidth almost 10 MHz) would be replaced by a data stream (bandwidth about 10 Kbaud) that drives a "talking head" modeled after each participant's facial appearance, and tracking his/her expression, lip movements and other gestures well enough to substitute for actual video, but at one-thousandth its bandwidth. Such applications unite vision and graphics, since model-based image understanding, tracking, and encoding are required at the transmitting end, capable of decoding into an articulated graphics emulation (avatar) of each participant at the receiving end. The MPEG-7 standard due for the year 2001 will extend these requirements to include provision for content-based video information retrieval, such as: "Find

me all video sequences in a digital library that show this actress blowing a kiss." Although media laboratories have tended to demonstrate tracking capabilities using cartoon characters and the development of more engaging video games (see Figs. 6a, and 6b of the "Pfinder" paper in this issue), potential applications of such environments for grown-ups are certain to emerge.

Another goal for the robust representation and understanding of facial information is computer recognition of personal identity. Besides its potential roles in office automation, log in control, and the automatic personalization of environments, obvious applications abound in security systems, criminology, physical access control, and prevention of fraud. Until today, humans have proved or verified their personal identity in their interactions with machines either through owning a special possession, such as a metal key or a plastic card, or through having secret knowledge such as a password or PIN number. (Metal keys date back 5,500 years to the Bronze Age, and passwords at least to Roman Centurions, yet today we still remain annoyingly burdened with too many of each.) But tests for mechanical objects like keys, or for secret knowledge like passwords, only indicate that they were present at the transaction, not that their rightful owners were. Face recognition and other forms of biometric identification seek to use a feature of the body or a unique behavior pattern of an individual to authorize a transaction, provided that the feature or the behavior pattern in question spans enough degrees-of-freedom of statistical variation across the human population to be considered unique for that individual.

It remains to be established just how unique facial appearance is, and whether algorithms can be designed that might approximate the face recognition performance of humans, for whom error rates are slightly (but not much) below one percent. One upper bound, but not the lowest one, for face recognition is set by the birth-rate of identical twins, which is also below one percent. Monozygotic twins are genetically identical (hence their birth rate also sets an upper confidence bound for DNA-based identification), and they illustrate the dramatic genetic penetrance of facial appearance. The main phenotypic factor (reflecting development, whereas a genotypic factor reflects the genome) for facial appearance is aging, but obviously this cannot be used as a basis for distinguishing identical twins. Nevertheless humans are normally able to distinguish between identical twins who are well-known to them, by picking up on any phenotypic differences. It remains to be seen whether computer vision algorithms can pass this same test. In any case, performance in face recognition at the 99 percent level, if that limit can be achieved, will not be good enough for high-security applications, nor for those

• John Daugman is with the Computer Laboratory, University of Cambridge, and he is an Associate Editor of PAMI.
E-mail: John.Daugman@CL.cam.ac.uk

that require exhaustive search through very large databases rather than merely a one-to-one verification match. Nonetheless, achievement of a 99 percent correct rate for automatic, passive, general-purpose face recognition would still find significant applications.

As it happens, there is unique phenotypic information of high complexity within faces that can be used as the basis for recognition of personal identity with very high confidence levels. This is the random texture that is visible in the eye's iris, which spans about 267 independent degrees-of-freedom of local phase variation and is stable over life. These random patterns behind the cornea are the largest source of accessible degrees-of-freedom (stable forms of variation across individuals) to be found in faces: greater in number by about an order of magnitude than those spanned by existing coding schemes for gross facial appearance. The "information density" of the iris, as defined by its human-population entropy per unit tissue area, works out to 3.4 bits per square millimeter. However, the small (1 cm) size of an iris requires that imaging be done at distances not exceeding about one meter, which is a limitation not suffered by schemes that encode the grosser degrees-of-freedom visible in the overall facial appearance.

The central factors determining the performance of face (or other) recognition systems are the dimensionality and the variance of the set of degrees-of-freedom which are encoded. Ideally their inter-class variance should be large and their intra-class variance should be small, so that different faces generate face codes that are as different as possible from each other, while different images of the same face generate very similar codes. Recent investigations of how well this goal is achieved have studied the invariances (or lack thereof) in face coding schemes under changes in illumination, perspective angle or pose, and expression. Their results have tended to show that there is greater variability in a given face across these three types of changes, than there is among different faces when these three factors are held constant. This is a fundamental problem in terms of classical pattern recognition theory, and several of the papers in this theme section address this topic.

A face is a surface of a three-dimensional solid having partially deformable parts. The images it projects depend upon pose, perspective angle, illumination conditions, age, cosmetics or adornments, and expression. A debate continues over whether it is better to represent faces with 2D appearance-based or 3D model-based methods. In either case, what are the optimal degrees-of-freedom to extract? Which ones are generic for all faces (and hence relevant for the face detection problem), and which ones are particular for a single face (and hence relevant for the recognition problem)? How should the integration of evidence be performed, and how should decisions under uncertainty best be made? In all of these respects, face recognition is both a "Holy-Grail" problem for machine intelligence, and also a problem that is paradigmatic for all of pattern recognition.

It is therefore all the more remarkable that primate brains have solved these problems. The neural mechanisms that evolved include highly specialized brain areas for face processing, as evidenced partly by the neurological syndrome called *facial prosopagnosia* following certain brain

traumas or stroke, in which vision remains normal in all respects except that the patient can no longer recognize faces, either as particular persons or even as a distinct class of visual object. In non-human primates, electrophysiological recordings from single neurons have revealed that more than 30 distinct subareas of the inferotemporal lobe of the brain are concerned with face processing, including neurons whose response properties are especially sensitive to the direction of gaze in the other's face and whether eye-contact is being made. (Among primates, both human and non-human, eye-contact is a very loaded signal.)

There is general consensus in behavioral biology today that the main factor driving the evolution of large brains in primates was the computational load of *social* processing. We evolved large brains not because of the reproductive advantages of being able to become computer scientists, but rather because large brains could better support our various social conspiracies: identifying, outwitting, controlling, enlisting, and frightening each other; charming, deceiving, seducing, betraying; knowing our place in a power hierarchy, especially in relation to the alpha male; and acquiring the ability to manipulate the behavior and mental states of each other as effectively as possible. Facial, postural, and gestural expression and identification were crucial elements in those social computations, and our practice of them today is but a recent gloss on ancient biological foundations. It will be interesting to see whether our efforts now in machine intelligence to build interfaces to these quintessential phenomena of human appearance, expression, and gesture will succeed to the point that we can interact with machines with all due ceremony as with ourselves.



John Daugman is a tenured faculty member at Cambridge University, where he teaches and conducts research in The Computer Laboratory. He received his degrees at Harvard University and taught on the faculty at Harvard from 1985 to 1989, when he became the inaugural holder of the Toshiba Endowed Chair in Computer Science at the Tokyo Institute of Technology. He was awarded a Presidential Young Investigator Award by the US National Science Foundation for his work in computational neuroscience and

computer vision.

One of his major research interests involves modeling biological strategies for visual processing, and trying to emulate these when possible in artificial systems. Among his technical contributions are the invention of iris recognition (IEEE PAMI-15, 1993), for which Patent 5,291,560 was awarded, and for which international licenses have been issued for several applications; and his development of the 2D Gabor Transform (IEEE ASSP-36, 1988) which today is widely applied across computer vision domains including face recognition, motion detection, texture classification, and optical character recognition. He has served for four years as an Associate Editor of PAMI.