Face Recognition: A Literature Survey

W. ZHAO

Sarnoff Corporation

R. CHELLAPPA

University of Maryland

P. J. PHILLIPS

National Institute of Standards and Technology

AND

A. ROSENFELD

University of Maryland

As one of the most successful applications of image analysis and understanding, face recognition has recently received significant attention, especially during the past several years. At least two reasons account for this trend: the first is the wide range of commercial and law enforcement applications, and the second is the availability of feasible technologies after 30 years of research. Even though current machine recognition systems have reached a certain level of maturity, their success is limited by the conditions imposed by many real applications. For example, recognition of face images acquired in an outdoor environment with changes in illumination and/or pose remains a largely unsolved problem. In other words, current systems are still far away from the capability of the human perception system.

This paper provides an up-to-date critical survey of still- and video-based face recognition research. There are two underlying motivations for us to write this survey paper: the first is to provide an up-to-date review of the existing literature, and the second is to offer some insights into the studies of machine recognition of faces. To provide a comprehensive survey, we not only categorize existing recognition techniques but also present detailed descriptions of representative methods within each category. In addition, relevant topics such as psychophysical studies, system evaluation, and issues of illumination and pose variation are covered.

Categories and Subject Descriptors: I.5.4 [**Pattern Recognition**]: Applications

General Terms: Algorithms

Additional Key Words and Phrases: Face recognition, person identification

An earlier version of this paper appeared as "Face Recognition: A Literature Survey," Technical Report CAR-TR-948, Center for Automation Research, University of Maryland, College Park, MD, 2000.

Authors' addresses: W. Zhao, Vision Technologies Lab, Sarnoff Corporation, Princeton, NJ 08543-5300; email: wzhao@sarnoff.com; R. Chellappa and A. Rosenfeld, Center for Automation Research, University of Maryland, College Park, MD 20742-3275; email: {rama,ar}@cfar.umd.edu; P. J. Phillips, National Institute of Standards and Technology, Gaithersburg, MD 20899; email: jonathon@nist.gov.

1. INTRODUCTION

As one of the most successful applications of image analysis and understanding, face recognition has recently received significant attention, especially during the past few years. This is evidenced by the emergence of face recognition conferences such as the International Conference on Audio- and Video-Based Authentication (AVBPA) since 1997 and the International Conference on Automatic Face and Gesture Recognition (AFGR) since 1995, systematic empirical evaluations of face recognition techniques (FRT), including the FERET [Phillips et al. 1998b, 2000; Rizvi et al. 1998], FRVT 2000 [Blackburn et al. 2001], FRVT 2002 [Phillips et al. 2003], and XM2VTS [Messer et al. 1999] protocols, and many commercially available systems (Table II). There are at least two reasons for this trend; the first is the wide range of commercial and law enforcement applications and the second is the availability of feasible technologies after 30 years of research. In addition, the problem of machine recognition of human faces continues to attract researchers from disciplines such as image processing, pattern recognition, neural networks, computer vision, computer graphics, and psychology.

The strong need for user-friendly systems that can secure our assets and protect our privacy without losing our identity in a sea of numbers is obvious. At present, one needs a PIN to get cash from an ATM, a password for a computer, a dozen others to access the internet, and so on. Although very reliable methods of biometric personal identification exist, for example, fingerprint analysis and retinal or iris scans, these methods rely on the cooperation of the participants, whereas a personal identification system based on analysis of frontal or profile images of the face is often effective without the participant's cooperation or knowledge. Some of the advantages/disadvantages of different biometrics are described in Phillips et al. [1998]. Table I lists some of the applications of face recognition.

Commercial and law enforcement applications of FRT range from static, controlled-format photographs to uncontrolled video images, posing a wide range of technical challenges and requiring an equally wide range of techniques from image processing, analysis, understanding, and pattern recognition. One can broadly classify FRT systems into two groups depending on whether they make use of static images or of video. Within these groups, significant differences exist, depending on the specific application. The differences are in terms of image quality, amount of background clutter (posing challenges to segmentation algorithms), variability of the images of a particular individual that must be recognized, availability of a well-defined recognition or matching criterion, and the nature, type, and amount of input from a user. A list of some commercial systems is given in Table II.

A general statement of the problem of machine recognition of faces can be formulated as follows: given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces. Available

Table I. Typical Applications of Face Recognition

Areas	Specific applications
Entertainment	Video game, virtual reality, training programs
	Human-robot-interaction, human-computer-interaction
Smart cards	Drivers' licenses, entitlement programs
	Immigration, national ID, passports, voter registration
	Welfare fraud
Information security	TV Parental control, personal device logon, desktop logon
	Application security, database security, file encryption
	Intranet security, internet access, medical records
	Secure trading terminals
Law enforcement and surveillance	Advanced video surveillance, CCTV control
	Portal control, postevent analysis
	Shoplifting, suspect tracking and investigation

Table II. Available Commercial Face Recognition Systems (Some of these Web sites may have changed or been removed.) [The identification of any company, commercial product, or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology or any of the authors or their institutions.]

Commercial products	Websites
FaceIt from Visionics	http://www.FaceIt.com
Viisage Technology	http://www.viisage.com
FaceVACS from Plettac	http://www.plettac-electronics.com
FaceKey Corp.	http://www.facekey.com
Cognitec Systems	http://www.cognitec-systems.de
Keyware Technologies	http://www.keywareusa.com/
Passfaces from ID-arts	http://www.id-arts.com/
ImageWare Sofware	http://www.iwsinc.com/
Eyematic Interfaces Inc.	http://www.eyematic.com/
BioID sensor fusion	http://www.bioid.com
Visionsphere Technologies	http://www.visionspheretech.com/menu.htm
Biometric Systems, Inc.	http://www.biometrica.com/
FaceSnap Recoder	http://www.facesnap.de/htdocs/english/index2.html
SpotIt for face composite	http://spotit.itc.it/SpotIt.html

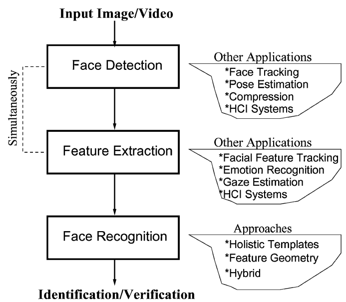

Fig. 1. Configuration of a generic face recognition system.

collateral information such as race, age, gender, facial expression, or speech may be used in narrowing the search (enhancing recognition). The solution to the problem involves segmentation of faces (face detection) from cluttered scenes, feature extraction from the face regions, recognition, or verification (Figure 1). In identification problems, the input to the system is an unknown face, and the system reports back the determined identity from a database of known individuals, whereas in verification problems, the system needs to confirm or reject the claimed identity of the input face.

Face perception is an important part of the capability of human perception system and is a routine task for humans, while building a similar computer system is still an on-going research area. The earliest work on face recognition can be traced back at least to the 1950s in psychology [Bruner and Tagiuri 1954] and to the 1960s in the engineering literature [Bledsoe 1964]. Some of the earliest studies include work on facial expression of emotions by Darwin [1972] (see also Ekman [1998]) and on facial profile-based biometrics by Galton [1888]). But research on automatic machine recognition of faces really started in the 1970s [Kelly 1970] and after the seminal work of Kanade [1973]. Over the past 30 years extensive research has been conducted by psychophysicists, neuroscientists, and engineers on various aspects of face recognition by humans and machines. Psychophysicists and neuroscientists have been concerned with issues such as whether face perception is a dedicated process (this issue is still being debated in the psychology community [Biederman and Kalocsai 1998; Ellis 1986; Gauthier et al. 1999; Gauthier and Logothetis 2000]) and whether it is done holistically or by local feature analysis.

Many of the hypotheses and theories put forward by researchers in these disciplines have been based on rather small sets of images. Nevertheless, many of the

findings have important consequences for engineers who design algorithms and systems for machine recognition of human faces. Section 2 will present a concise review of these findings.

Barring a few exceptions that use range data [Gordon 1991], the face recognition problem has been formulated as recognizing three-dimensional (3D) objects from two-dimensional (2D) images.[1] Earlier approaches treated it as a 2D pattern recognition problem. As a result, during the early and mid-1970s, typical pattern classification techniques, which use measured attributes of features (e.g., the distances between important points) in faces or face profiles, were used [Bledsoe 1964; Kanade 1973; Kelly 1970]. During the 1980s, work on face recognition remained largely dormant. Since the early 1990s, research interest in FRT has grown significantly. One can attribute this to several reasons: an increase in interest in commercial opportunities; the availability of real-time hardware; and the increasing importance of surveillance-related applications.

Over the past 15 years, research has focused on how to make face recognition systems fully automatic by tackling problems such as localization of a face in a given image or video clip and extraction of features such as eyes, mouth, etc. Meanwhile, significant advances have been made in the design of classifiers for successful face recognition. Among appearance-based holistic approaches, eigenfaces [Kirby and Sirovich 1990; Turk and Pentland 1991] and Fisherfaces [Belhumeur et al. 1997; Etemad and Chellappa 1997; Zhao et al. 1998] have proved to be effective in experiments with large databases. Feature-based graph matching approaches [Wiskott et al. 1997] have also been quite successful. Compared to holistic approaches, feature-based methods are less sensitive to variations in illumination and viewpoint and to inaccuracy in face localization. However, the feature extraction techniques needed for this type of approach are still not reliable or accurate enough [Cox et al. 1996]. For example, most eye localization techniques assume some geometric and textural models and do not work if the eye is closed. Section 3 will present a review of still-image-based face recognition.

During the past 5 to 8 years, much research has been concentrated on video-based face recognition. The still image problem has several inherent advantages and disadvantages. For applications such as drivers' licenses, due to the controlled nature of the image acquisition process, the segmentation problem is rather easy. However, if only a static picture of an airport scene is available, automatic location and segmentation of a face could pose serious challenges to any segmentation algorithm. On the other hand, if a video sequence is available, segmentation of a moving person can be more easily accomplished using motion as a cue. But the small size and low image quality of faces captured from video can significantly increase the difficulty in recognition. Video-based face recognition is reviewed in Section 4.

As we propose new algorithms and build more systems, measuring the performance of new systems and of existing systems becomes very important. Systematic data collection and evaualation of face recognition systems is reviewed in Section 5.

Recognizing a 3D object from its 2D images poses many challenges. The illumination and pose problems are two prominent issues for appearance- or image-based approaches. Many approaches have been proposed to handle these issues, with the majority of them exploring domain knowledge. Details of these approaches are discussed in Section 6.

In 1995, a review paper [Chellappa et al. 1995] gave a thorough survey of FRT at that time. (An earlier survey [Samal and Iyengar 1992] appeared in 1992.) At that time, video-based face recognition was still in a nascent stage. During the past 8 years, face recognition has received increased attention and has advanced

[1]There have been recent advances on 3D face recognition in situations where range data acquired through structured light can be matched reliably [Bronstein et al. 2003].

technically. Many commercial systems for still face recognition are now available. Recently, significant research efforts have been focused on video-based face modeling/tracking, recognition, and system integration. New datasets have been created and evaluations of recognition techniques using these databases have been carried out. It is not an overstatement to say that face recognition has become one of the most active applications of pattern recognition, image analysis and understanding.

In this paper we provide a critical review of current developments in face recognition. This paper is organized as follows: in Section 2 we briefly review issues that are relevant from a psychophysical point of view. Section 3 provides a detailed review of recent developments in face recognition techniques using still images. In Section 4 face recognition techniques based on video are reviewed. Data collection and performance evaluation of face recognition algorithms are addressed in Section 5 with descriptions of representative protocols. In Section 6 we discuss two important problems in face recognition that can be mathematically studied, lack of robustness to illumination and pose variations, and we review proposed methods of overcoming these limitations. Finally, a summary and conclusions are presented in Section 7.

2. PSYCHOPHYSICS/NEUROSCIENCE ISSUES RELEVANT TO FACE RECOGNITION

Human recognition processes utilize a broad spectrum of stimuli, obtained from many, if not all, of the senses (visual, auditory, olfactory, tactile, etc.). In many situations, contextual knowledge is also applied, for example, surroundings play an important role in recognizing faces in relation to where they are supposed to be located. It is futile to even attempt to develop a system using existing technology, which will mimic the remarkable face recognition ability of humans. However, the human brain has its limitations in the total number of persons that it can accurately "remember." A key advantage of a computer system is its capacity to handle large numbers of face images. In most applications the images are available only in the form of single or multiple views of 2D intensity data, so that the inputs to computer face recognition algorithms are visual only. For this reason, the literature reviewed in this section is restricted to studies of human visual perception of faces.

Many studies in psychology and neuroscience have direct relevance to engineers interested in designing algorithms or systems for machine recognition of faces. For example, findings in psychology [Bruce 1988; Shepherd et al. 1981] about the relative importance of different facial features have been noted in the engineering literature [Etemad and Chellappa 1997]. On the other hand, machine systems provide tools for conducting studies in psychology and neuroscience [Hancock et al. 1998; Kalocsai et al. 1998]. For example, a possible engineering explanation of the bottom lighting effects studied in Johnston et al. [1992] is as follows: when the actual lighting direction is opposite to the usually assumed direction, a shape-from-shading algorithm recovers incorrect structural information and hence makes recognition of faces harder.

A detailed review of relevant studies in psychophysics and neuroscience is beyond the scope of this paper. We only summarize findings that are potentially relevant to the design of face recognition systems. For details the reader is referred to the papers cited below. Issues that are of potential interest to designers are[2]:

—*Is face recognition a dedicated process?* [Biederman and Kalocsai 1998; Ellis 1986; Gauthier et al. 1999; Gauthier and Logothetis 2000]: It is traditionally believed that face recognition is a dedicated process different from other object recognition tasks. Evidence for the existence of a dedicated face processing system comes from several sources [Ellis 1986]. (a) Faces are more easily remembered by humans than other objects when presented in an upright orientation. (b) Prosopagnosia patients are unable to recognize previously familiar faces, but usually have no other profound agnosia. They recognize people by their voices, hair color, dress, etc. It should be noted that prosopagnosia patients recognize whether a given object is a face or not, but then have difficulty in identifying the face. Seven differences between face recognition and object recognition can be summarized [Biederman and Kalocsai 1998] based on empirical evidence: (1) *configural effects* (related to the choice of different types of machine recognition systems), (2) *expertise*, (3) *differences verbalizable*, (4) *sensitivity to contrast polarity and illumination direction* (related to the illumination problem in machine recognition systems), (5) *metric variation*, (6) *Rotation in depth* (related to the pose variation problem in machine recognition systems), and (7) *rotation in plane/inverted face*. Contrary to the traditionally held belief, some recent findings in human neuropsychology and neuroimaging suggest that face recognition may not be unique. According to [Gauthier and Logothetis 2000], recent neuroimaging studies in humans indicate that level of categorization and expertise interact to produce the specification for faces in the middle fusiform gyrus.[3] Hence it is possible that the encoding scheme used for faces may also be employed for other classes with similar properties. (On recognition of familiar vs. unfamiliar faces see Section 7.)

—*Is face perception the result of holistic or feature analysis?* [Bruce 1988; Bruce et al. 1998]: Both holistic and feature information are crucial for the perception and recognition of faces. Studies suggest the possibility of global descriptions serving as a front end for finer, feature-based perception. If dominant features are present, holistic descriptions may not be used. For example, in face recall studies, humans quickly focus on odd features such as big ears, a crooked nose, a staring eye, etc. One of the strongest pieces of evidence to support the view that face recognition involves more configural/holistic processing than other object recognition has been the face inversion effect in which an inverted face is much harder to recognize than a normal face (first demonstrated in [Yin 1969]). An excellent example is given in [Bartlett and Searcy 1993] using the "Thatcher illusion" [Thompson 1980]. In this illusion, the eyes and mouth of an expressing face are excised and inverted, and the result looks grotesque in an upright face; however, when shown inverted, the face looks fairly normal in appearance, and the inversion of the internal features is not readily noticed.

—*Ranking of significance of facial features* [Bruce 1988; Shepherd et al. 1981]: Hair, face outline, eyes, and mouth (not necessarily in this order) have been determined to be important for perceiving and remembering faces [Shepherd et al. 1981]. Several studies have shown that the nose plays an insignificant role; this may be due to the fact that almost all of these studies have been done using frontal images. In face recognition using profiles (which may be important in mugshot matching applications, where profiles can be extracted from side views), a distinctive nose shape could be more important than the eyes or mouth [Bruce 1988]. Another outcome of some studies is that both external and internal features are important in the recognition of previously presented but otherwise unfamiliar faces, but internal features are more dominant in the recognition of familiar faces. It has also been found that the upper part of the face is more useful for face recognition than the lower part [Shepherd et al. 1981]. The role of aesthetic attributes such as beauty, attractiveness, and/or pleasantness has also been studied, with the conclusion that

[2]Readers should be aware of the existence of diverse opinions on some of these issues. The opinions given here do not necessarily represent our views.

[3]The fusiform gyrus or occipitotemporal gyrus, located on the ventromedial surface of the temporal and occipital lobes, is thought to be critical for face recognition.

the more attractive the faces are, the better is their recognition rate; the least attractive faces come next, followed by the midrange faces, in terms of ease of being recognized.

—*Caricatures* [Brennan 1985; Bruce 1988; Perkins 1975]: A caricature can be formally defined [Perkins 1975] as "a symbol that exaggerates measurements relative to any measure which varies from one person to another." Thus the length of a nose is a measure that varies from person to person, and could be useful as a symbol in caricaturing someone, but not the number of ears. A standard caricature algorithm [Brennan 1985] can be applied to different qualities of image data (line drawings and photographs). Caricatures of line drawings do not contain as much information as photographs, but they manage to capture the important characteristics of a face; experiments based on nonordinary faces comparing the usefulness of line-drawing caricatures and unexaggerated line drawings decidedly favor the former [Bruce 1988].

—*Distinctiveness* [Bruce et al. 1994]: Studies show that distinctive faces are better retained in memory and are recognized better and faster than typical faces. However, if a decision has to be made as to whether an object is a face or not, it takes longer to recognize an atypical face than a typical face. This may be explained by different mechanisms being used for detection and for identification.

—*The role of spatial frequency analysis* [Ginsburg 1978; Harmon 1973; Sergent 1986]: Earlier studies [Ginsburg 1978; Harmon 1973] concluded that information in low spatial frequency bands plays a dominant role in face recognition. Recent studies [Sergent 1986] have shown that, depending on the specific recognition task, the low, bandpass and high-frequency components may play different roles. For example gender classification can be successfully accomplished using low-frequency components only, while identification requires the use of high-frequency components [Sergent 1986]. Low-frequency components contribute to global description, while high-frequency components contribute to the finer details needed in identification.

—*Viewpoint-invariant recognition?* [Biederman 1987; Hill et al. 1997; Tarr and Bulthoff 1995]: Much work in visual object recognition (e.g. [Biederman 1987]) has been cast within a theoretical framework introduced in [Marr 1982] in which different views of objects are analyzed in a way which allows access to (largely) viewpoint-invariant descriptions. Recently, there has been some debate about whether object recognition is viewpoint-invariant or not [Tarr and Bulthoff 1995]. Some experiments suggest that memory for faces is highly viewpoint-dependent. Generalization even from one profile viewpoint to another is poor, though generalization from one three-quarter view to the other is very good [Hill et al. 1997].

—*Effect of lighting change* [Bruce et al. 1998; Hill and Bruce 1996; Johnston et al. 1992]: It has long been informally observed that photographic negatives of faces are difficult to recognize. However, relatively little work has explored why it is so difficult to recognize negative images of faces. In [Johnston et al. 1992], experiments were conducted to explore whether difficulties with negative images and inverted images of faces arise because each of these manipulations reverses the apparent direction of lighting, rendering a top-lit image of a face apparently lit from below. It was demonstrated in [Johnston et al. 1992] that bottom lighting does indeed make it harder to identity familiar faces. In [Hill and Bruce 1996], the importance of top lighting for face recognition was demonstrated using a different task: matching surface images of faces to determine whether they were identical.

—*Movement and face recognition* [O'Toole et al. 2002; Bruce et al. 1998; Knight and Johnston 1997]: A recent study [Knight and Johnston 1997] showed that famous faces are easier to recognize when shown in moving sequences than in still photographs. This observation has been extended to show that movement helps in the recognition of familiar faces shown under a range of different types of degradations—negated, inverted, or thresholded [Bruce et al. 1998]. Even more interesting is the observation that there seems to be a benefit due to movement even if the information content is equated in the moving and static comparison conditions. However, experiments with unfamiliar faces suggest no additional benefit from viewing animated rather than static sequences.

—*Facial expressions* [Bruce 1988]: Based on neurophysiological studies, it seems that analysis of facial expressions is accomplished in parallel to face recognition. Some prosopagnosic patients, who have difficulties in identifying familiar faces, nevertheless seem to recognize expressions due to emotions. Patients who suffer from "organic brain syndrome" suffer from poor expression analysis but perform face recognition quite well.[4] Similarly, separation of face recognition and "focused visual processing" tasks (e.g., looking for someone with a thick mustache) have been claimed.

3. FACE RECOGNITION FROM STILL IMAGES

As illustrated in Figure 1, the problem of automatic face recognition involves three key steps/subtasks: (1) detection and rough normalization of faces, (2) feature extraction and accurate normalization of faces, (3) identification and/or verification. Sometimes, different subtasks are not totally separated. For example, the facial features (eyes, nose, mouth) used for face recognition are often used in face detection. Face detection and feature extraction can be achieved simultaneously, as indicated in Figure 1. Depending on the nature of the application, for example, the sizes of the training and testing databases, clutter and variability of the background, noise, occlusion, and speed requirements, some of the subtasks can be very challenging.

Though fully automatic face recognition systems must perform all three subtasks, research on each subtask is critical. This is not only because the techniques used for the individual subtasks need to be improved, but also because they are critical in many different applications (Figure 1). For example, face detection is needed to initialize face tracking, and extraction of facial features is needed for recognizing human emotion, which is in turn essential in human-computer interaction (HCI) systems. Isolating the subtasks makes it easier to assess and advance the state of the art of the component techniques. Earlier face detection techniques could only handle single or a few well-separated frontal faces in images with simple backgrounds, while state-of-the-art algorithms can detect faces and their poses in cluttered backgrounds [Gu et al. 2001; Heisele et al. 2001; Schneiderman and Kanade 2000; Viola and Jones 2001]. Extensive research on the subtasks has been carried out and relevant surveys have appeared on, for example, the subtask of face detection [Hjelmas and Low 2001; Yang et al. 2002].

In this section we survey the state of the art of face recognition in the engineering literature. For the sake of completeness, in Section 3.1 we provide a highlighted summary of research on face segmentation/detection and feature extraction. Section 3.2 contains detailed reviews of recent work on intensity image-based face recognition and categorizes methods of recognition from intensity images. Section 3.3 summarizes the status of face recognition and discusses open research issues.

3.1. Key Steps Prior to Recognition: Face Detection and Feature Extraction

The first step in any automatic face recognition systems is the detection of faces in images. Here we only provide a summary on this topic and highlight a few

[4]From a machine recognition point of view, dramatic facial expressions may affect face recognition performance if only one photograph is available.

very recent methods. After a face has been detected, the task of feature extraction is to obtain features that are fed into a face classification system. Depending on the type of classification system, features can be local features such as lines or fiducial points, or facial features such as eyes, nose, and mouth. Face detection may also employ features, in which case features are extracted simultaneously with face detection. Feature extraction is also a key to animation and recognition of facial expressions.

Without considering feature locations, face detection is declared successful if the presence and rough location of a face has been correctly identified. However, without accurate face and feature location, noticeable degradation in recognition performance is observed [Martinez 2002; Zhao 1999]. The close relationship between feature extraction and face recognition motivates us to review a few feature extraction methods that are used in the recognition approaches to be reviewed in Section 3.2. Hence, this section also serves as an introduction to the next section.

3.1.1. Segmentation/Detection: Summary. Up to the mid-1990s, most work on segmentation was focused on single-face segmentation from a simple or complex background. These approaches included using a whole-face template, a deformable feature-based template, skin color, and a neural network.

Significant advances have been made in recent years in achieving automatic face detection under various conditions. Compared to feature-based methods and template-matching methods, appearance- or image-based methods [Rowley et al. 1998; Sung and Poggio 1997] that train machine systems on large numbers of samples have achieved the best results. This may not be surprising since face objects are complicated, very similar to each other, and different from nonface objects. Through extensive training, computers can be quite good at detecting faces.

More recently, detection of faces under rotation in depth has been studied. One approach is based on training on multiple-view samples [Gu et al. 2001; Schneiderman and Kanade 2000]. Compared to invariant-feature-based methods [Wiskott et al. 1997], multiview-based methods of face detection and recognition seem to be able to achieve better results when the angle of out-of-plane rotation is large (35°). In the psychology community, a similar debate exists on whether face recognition is viewpoint-invariant or not. Studies in both disciplines seem to support the idea that for small angles, face perception is view-independent, while for large angles, it is view-dependent.

In a detection problem, two statistics are important: true positives (also referred to as *detection rate*) and false positives (reported detections in nonface regions). An ideal system would have very high true positive and very low false positive rates. In practice, these two requirements are conflicting. Treating face detection as a two-class classification problem helps to reduce false positives dramatically [Rowley et al. 1998; Sung and Poggio 1997] while maintaining true positives. This is achieved by retraining systems with false-positive samples that are generated by previously trained systems.

3.1.2. Feature Extraction: Summary and Methods

3.1.2.1. Summary. The importance of facial features for face recognition cannot be overstated. Many face recognition systems need facial features in addition to the holistic face, as suggested by studies in psychology. It is well known that even holistic matching methods, for example, eigenfaces [Turk and Pentland 1991] and Fisherfaces [Belhumeur et al. 1997], need accurate locations of key facial features such as eyes, nose, and mouth to normalize the detected face [Martinez 2002; Yang et al. 2002].

Three types of feature extraction methods can be distinguished: (1) generic methods based on edges, lines, and curves; (2) feature-template-based methods that are used to detect facial features such as eyes; (3) structural matching methods

that take into consideration geometrical constraints on the features. Early approaches focused on individual features; for example, a template-based approach was described in [Hallinan 1991] to detect and recognize the human eye in a frontal face. These methods have difficulty when the appearances of the features change significantly, for example, closed eyes, eyes with glasses, open mouth. To detect the features more reliably, recent approaches have used structural matching methods, for example, the Active Shape Model [Cootes et al. 1995]. Compared to earlier methods, these recent statistical methods are much more robust in terms of handling variations in image intensity and feature shape.

An even more challenging situation for feature extraction is feature "restoration," which tries to recover features that are invisible due to large variations in head pose. The best solution here might be to hallucinate the missing features either by using the bilateral symmetry of the face or using learned information. For example, a view-based statistical method claims to be able to handle even profile views in which many local features are invisible [Cootes et al. 2000].

3.1.2.2. Methods. A template-based approach to detecting the eyes and mouth in real images was presented in [Yuille et al. 1992]. This method is based on matching a predefined parameterized template to an image that contains a face region. Two templates are used for matching the eyes and mouth respectively. An energy function is defined that links edges, peaks and valleys in the image intensity to the corresponding properties in the template, and this energy function is minimized by iteratively changing the parameters of the template to fit the image. Compared to this model, which is manually designed, the *statistical* shape model (Active Shape Model, ASM) proposed in [Cootes et al. 1995] offers more flexibility and robustness. The advantages of using the so-called analysis through synthesis approach come from the fact that the solution is *constrained* by a *flexible* statistical model. To account for texture variation, the ASM model has been expanded to statistical appearance models including a Flexible Appearance Model (FAM) [Lanitis et al. 1995] and an Active Appearance Model (AAM) [Cootes et al. 2001]. In [Cootes et al. 2001], the proposed AAM combined a model of shape variation (i.e., ASM) with a model of the appearance variation of shape-normalized (shape-free) textures. A training set of 400 images of faces, each manually labeled with 68 landmark points, and approximately 10,000 intensity values sampled from facial regions were used. The shape model (mean shape, orthogonal mapping matrix $\mathbf{P}_s$ and projection vector $\mathbf{b}_s$) is generated by representing each set of landmarks as a vector and applying principal-component analysis (PCA) to the data. Then, after each sample image is warped so that its landmarks match the mean shape, texture information can be sampled from this shape-free face patch. Applying PCA to this data leads to a shape-free texture model (mean texture, $\mathbf{P}_g$ and $\mathbf{b}_g$). To explore the correlation between the shape and texture variations, a third PCA is applied to the concatenated vectors ($\mathbf{b}_s$ and $\mathbf{b}_g$) to obtain the combined model in which one vector $\mathbf{c}$ of appearance parameters controls both the shape and texture of the model. To match a given image and the model, an optimal vector of parameters (displacement parameters between the face region and the model, parameters for linear intensity adjustment, and the appearance parameters $\mathbf{c}$) are searched by minimizing the difference between the synthetic image and the given one. After matching, a best-fitting model is constructed that gives the locations of all the facial features and can be used to reconstruct the original images. Figure 2 illustrates the optimization/search procedure for fitting the model to the image. To speed up the search procedure, an efficient method is proposed that exploits the similarities among optimizations. This allows the direct method to find and apply directions of rapid convergence which are learned off-line.

Fig. 2. Multiresolution search from a displaced position using a face model. (Courtesy of T. Cootes, K. Walker, and C. Taylor.)

3.2. Recognition from Intensity Images

Many methods of face recognition have been proposed during the past 30 years. Face recognition is such a challenging yet interesting problem that it has attracted researchers who have different backgrounds: psychology, pattern recognition, neural networks, computer vision, and computer graphics. It is due to this fact that the literature on face recognition is vast and diverse. Often, a single system involves techniques motivated by different principles. The usage of a mixture of techniques makes it difficult to classify these systems based purely on what types of techniques they use for feature representation or classification. To have a clear and high-level categorization, we instead follow a guideline suggested by the psychological study of how humans use holistic and local features. Specifically, we have the following categorization:

(1) *Holistic matching methods.* These methods use the whole face region as the raw input to a recognition system. One of the most widely used representations of the face region is eigenpictures [Kirby and Sirovich 1990; Sirovich and Kirby 1987], which are based on principal component analysis.

(2) *Feature-based (structural) matching methods.* Typically, in these methods, local features such as the eyes, nose, and mouth are first extracted and their locations and local statistics (geometric and/or appearance) are fed into a structural classifier.

(3) *Hybrid methods.* Just as the human perception system uses both local features and the whole face region to recognize a face, a machine recognition system should use both. One can argue that these methods could potentially offer the best of the two types of methods.

Within each of these categories, further classification is possible (Table III). Using principal-component analysis (PCA), many face recognition techniques have been developed: eigenfaces [Turk and Pentland 1991], which use a nearest-neighbor classifier; feature-line-based methods, which replace the point-to-point distance with the distance between a point and the feature line linking two stored sample points [Li and Lu 1999]; Fisherfaces [Belhumeur et al. 1997; Liu and Wechsler 2001; Swets and Weng 1996b; Zhao et al. 1998] which use linear/Fisher discriminant analysis (FLD/LDA) [Fisher 1938]; Bayesian methods, which use a probabilistic distance metric [Moghaddam and Pentland 1997]; and SVM methods, which use a support vector machine as the classifier [Phillips 1998]. Utilizing higher-order statistics, independent-component

Table III. Categorization of Still Face Recognition Techniques

Approach	Representative work
Holistic methods	
Principal-component analysis (PCA)	
Eigenfaces	Direct application of PCA [Craw and Cameron 1996; Kirby and Sirovich 1990; Turk and Pentland 1991]
Probabilistic eigenfaces	Two-class problem with prob. measure [Moghaddam and Pentland 1997]
Fisherfaces/subspace LDA	FLD on eigenspace [Belhumeur et al. 1997; Swets and Weng 1996b; Zhao et al. 1998]
SVM	Two-class problem based on SVM [Phillips 1998]
Evolution pursuit	Enhanced GA learning [Liu and Wechsler 2000a]
Feature lines	Point-to-line distance based [Li and Lu 1999]
ICA	ICA-based feature analysis [Bartlett et al. 1998]
Other representations	
LDA/FLD	LDA/FLD on raw image [Etemad and Chellappa 1997]
PDBNN	Probabilistic decision based NN [Lin et al. 1997]
Feature-based methods	
Pure geometry methods	Earlier methods [Kanade 1973; Kelly 1970]; recent methods [Cox et al. 1996; Manjunath et al. 1992]
Dynamic link architecture	Graph matching methods [Okada et al. 1998; Wiskott et al. 1997]
Hidden Markov model	HMM methods [Nefian and Hayes 1998; Samaria 1994; Samaria and Young 1994]
Convolution Neural Network	SOM learning based CNN methods [Lawrence et al. 1997]
Hybrid methods	
Modular eigenfaces	Eigenfaces and eigenmodules [Pentland et al. 1994]
Hybrid LFA	Local feature method [Penev and Atick 1996]
Shape-normalized	Flexible appearance models [Lanitis et al. 1995]
Component-based	Face region and components [Huang et al. 2003]

analysis (ICA) is argued to have more representative power than PCA, and hence may provide better recognition performance than PCA [Bartlett et al. 1998]. Being able to offer potentially greater generalization through learning, neural networks/learning methods have also been applied to face recognition. One example is the Probabilistic Decision-Based Neural Network (PDBNN) method [Lin et al. 1997] and the other is the evolution pursuit (EP) method [Liu and Wechsler 2000a].

Most earlier methods belong to the category of structural matching methods, using the width of the head, the distances between the eyes and from the eyes to the mouth, etc. [Kelly 1970], or the distances and angles between eye corners, mouth extrema, nostrils, and chin top [Kanade 1973]. More recently, a mixture-distance based approach using manually extracted distances was reported [Cox et al. 1996]. Without finding the exact locations of facial features, Hidden Markov Model- (HMM-) based methods use strips of pixels that cover the forehead, eye, nose, mouth, and chin [Nefian and Hayes 1998; Samaria 1994; Samaria and Young 1994]. [Nefian and Hayes 1998] reported better performance than Samaria [1994] by using the KL projection coefficients instead of the strips of raw pixels. One of the most successful systems in this category is the graph matching system [Okada et al. 1998; Wiskott et al. 1997], which is based on the Dynamic Link Architecture (DLA) [Buhmann et al. 1990; Lades et al. 1993]. Using an unsupervised learning method based on a self-organizing map (SOM), a system based on a convolutional neural network (CNN) has been developed [Lawrence et al. 1997].

In the hybrid method category, we will briefly review the modular eigenface method [Pentland et al. 1994], a hybrid representation based on PCA and local feature analysis (LFA) [Penev and Atick 1996], a flexible appearance model-based method [Lanitis et al. 1995], and a recent development [Huang et al. 2003] along this direction. In [Pentland et al. 1994],

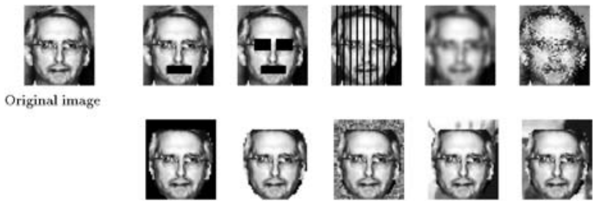

Fig. 3. Electronically modified images which were correctly identified.

the use of hybrid features by combining eigenfaces and other eigenmodules is explored: eigeneyes, eigenmouth, and eigennose. Though experiments show slight improvements over holistic eigenfaces or eigenmodules based on structural matching, we believe that these types of methods are important and deserve further investigation. Perhaps many relevant problems need to be solved before fruitful results can be expected, for example, how to optimally arbitrate the use of holistic and local features.

Many types of systems have been successfully applied to the task of face recognition, but they all have some advantages and disadvantages. Appropriate schemes should be chosen based on the specific requirements of a given task. Most of the systems reviewed here focus on the subtask of recognition, but others also include automatic face detection and feature extraction, making them fully automatic systems [Lin et al. 1997; Moghaddam and Pentland 1997; Wiskott et al. 1997].

3.2.1. Holistic Approaches

3.2.1.1. Principal-Component Analysis. Starting from the successful low-dimensional reconstruction of faces using KL or PCA projections [Kirby and Sirovich 1990; Sirovich and Kirby 1987], eigenpictures have been one of the major driving forces behind face representation, detection, and recognition. It is well known that there exist significant statistical redundancies in natural images [Ruderman 1994]. For a limited class of objects such as face images that are normalized with respect to scale, translation, and rotation, the redundancy is even greater [Penev and Atick 1996; Zhao 1999]. One of the best global compact representations is KL/PCA, which decorrelates the outputs. More specifically, sample vectors $\mathbf{x}$ can be expressed as linear combinations of the orthogonal basis Φ_i: $\mathbf{x} = \sum_{i=1}^{n} a_i \Phi_i \approx \sum_{i=1}^{m} a_i \Phi_i$ (typically $m \ll n$) by solving the eigenproblem

$$C\Phi = \Phi\Lambda, \tag{1}$$

where C is the covariance matrix for input $\mathbf{x}$.

An advantage of using such representations is their reduced sensitivity to noise. Some of this noise may be due to small occlusions, as long as the topological structure does not change. For example, good performance under blurring, partial occlusion and changes in background has been demonstrated in many eigenpicture-based systems, as illustrated in Figure 3. This should not come as a surprise, since the PCA reconstructed images are much better than the original distorted images in terms of their global appearance (Figure 4).

For better approximation of face images outside the training set, using an extended training set that adds mirror-imaged faces was shown to achieve lower approximation error [Kirby and Sirovich 1990]. Using such an extended training set, the eigenpictures are either symmetric or antisymmetric, with the most leading eigenpictures typically being symmetric.

Fig. 4. Reconstructed images using 300 PCA projection coefficients for electronically modified images (Figure 3). (From Zhao [1999].)

The first really successful demonstration of machine recognition of faces was made in [Turk and Pentland 1991] using eigenpictures (also known as eigenfaces) for face detection and identification. Given the eigenfaces, every face in the database can be represented as a vector of weights; the weights are obtained by projecting the image into eigenface components by a simple inner product operation. When a new test image whose identification is required is given, the new image is also represented by its vector of weights. The identification of the test image is done by locating the image in the database whose weights are the closest to the weights of the test image. By using the observation that the projection of a face image and a nonface image are usually different, a method of detecting the presence of a face in a given image is obtained. The method was demonstrated using a database of 2500 face images of 16 subjects, in all combinations of three head orientations, three head sizes, and three lighting conditions.

Using a probabilistic measure of similarity, instead of the simple Euclidean distance used with eigenfaces [Turk and Pentland 1991], the standard eigenface approach was extended [Moghaddam and Pentland 1997] to a Bayesian approach. Practically, the major drawback of a Bayesian method is the need to estimate probability distributions in a high-dimensional space from very limited numbers of training samples per class. To avoid this problem, a much simpler two-class problem was created from the multiclass problem by using a similarity measure based on a Bayesian analysis of image differences. Two mutually exclusive classes were defined: Ω_I, representing *intrapersonal* variations between multiple images of the same individual, and Ω_E, representing *extrapersonal* variations due to differences in identity. Assuming that both classes are Gaussian-distributed, likelihood functions $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ were estimated for a given intensity difference $\Delta = I_1 - I_2$. Given these likelihood functions and using the MAP rule, two face images are determined to belong to the same individual if $P(\Delta|\Omega_I) > P(\Delta|\Omega_E)$. A large performance improvement of this probabilistic matching technique over standard nearest-neighbor eigenspace matching was reported using large face datasets including the FERET database [Phillips et al. 2000]. In Moghaddam and Pentland [1997], an efficient technique of probability density estimation was proposed by decomposing the input space into two mutually exclusive subspaces: the principal subspace F and its orthogonal subspace $\hat{F}$ (a similar idea was explored in Sung and Poggio [1997]). Covariances only in the principal subspace are estimated for use in the Mahalanobis distance [Fukunaga 1989]. Experimental results have been reported using different subspace dimensionalities M_I and M_E for Ω_I and Ω_E. For example, $M_I = 10$ and $M_E = 30$ were used for internal tests, while $M_I = M_E = 125$ were used for the FERET test. In Figure 5, the so-called dual eigenfaces separately trained on samples from Ω_I and Ω_E are plotted along with the standard eigenfaces. While the extrapersonal

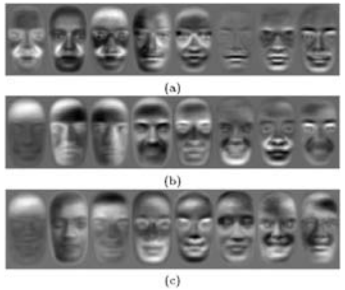

Fig. 5. Comparison of "dual" eigenfaces and standard eigenfaces: (a) intrapersonal, (b) extrapersonal, (c) standard [Moghaddam and Pentland 1997]. (Courtesy of B. Moghaddam and A. Pentland.)

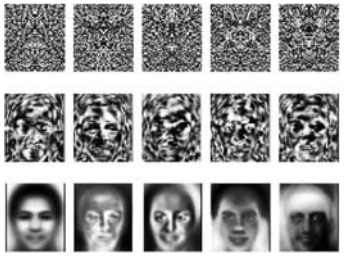

Fig. 6. Different projection bases constructed from a set of 444 individuals, where the set is augumented via adding noise and mirroring. The first row shows the first five pure LDA basis images W; the second row shows the first five subspace LDA basis images $W\Phi$; the average face and first four eigenfaces Φ are shown on the third row [Zhao et al. 1998].

eigenfaces appear more similar to the standard eigenfaces than the intrapersonal ones, the intrapersonal eigenfaces represent subtle variations due mostly to expression and lighting, suggesting that they are more critical for identification [Moghaddam and Pentland 1997].

Face recognition systems using LDA/FLD have also been very successful [Belhumeur et al. 1997; Etemad and Chellappa 1997; Swets and Weng 1996b; Zhao et al. 1998; Zhao et al. 1999]. LDA training is carried out via scatter matrix analysis [Fukunaga 1989]. For an M-class problem, the within- and between-class scatter matrices S_w, S_b are computed as follows:

$$S_w = \sum_{i=1}^{M} Pr(\omega_i) C_i, \qquad S_b = \sum_{i=1}^{M} Pr(\omega_i)(\mathbf{m}_i - \mathbf{m}_0)(\mathbf{m}_i - \mathbf{m}_0)^T, \tag{2}$$

where $Pr(\omega_i)$ is the prior class probability, and is usually replaced by $1/M$ in practice with the assumption of equal priors. Here S_w is the *within-class satter matrix*, showing the average scatter[5] C_i of the sample vectors $\mathbf{x}$ of different classes ω_i around their respective means $\mathbf{m}_i$: $C_i = E[(\mathbf{x}(\omega) - \mathbf{m}_i)(\mathbf{x}(\omega) - \mathbf{m}_i)^T | \omega = \omega_i]$. Similarly, S_b is the Between-class Scatter Matrix, representing the scatter of the conditional mean vectors $\mathbf{m}_i$ around the overall mean vector $\mathbf{m}_0$. A commonly used measure for quantifying discriminatory power is the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix: $\mathcal{J}(T) = |T^T S_b T| / |T^T S_w T|$. The optimal projection matrix W which maximizes $\mathcal{J}(T)$ can be obtained by solving a generalized eigenvalue problem:

$$S_b W = S_w W \Lambda_W. \tag{3}$$

It is helpful to make comparisons among the so-called (linear) projection algorithms. Here we illustrate the comparison between eigenfaces and Fisherfaces. Similar comparisons can be made for other methods, for example, ICA projection methods. In all these projection algorithms, classification is performed by (1) projecting the input $\mathbf{x}$ into a subspace via a projection/basis matrix $\mathbf{P}_{roj}$[6]:

[5]These are also conditional covariance matrices; the total covariance C used to compute the PCA projection is $C = \sum_{i=1}^{M} Pr(\omega_i) C_i$.

[6]$\mathbf{P}_{roj}$ is Φ for eigenfaces, W for Fisherfaces with pure LDA projection, and $W\Phi$ for Fisherfaces with sequential PCA and LDA projections; these three bases are shown for visual comparison in Figure 6.

$$\mathbf{z} = \mathbf{P}_{roj}\mathbf{x}; \tag{4}$$

(2) comparing the projection coefficient vector $\mathbf{z}$ of the input to all the prestored projection vectors of labeled classes to determine the input class label. The vector comparison varies in different implementations and can influence the system's performance dramatically [Moon and Phillips 2001]. For example, PCA algorithms can use either the angle or the Euclidean distance (weighted or unweighted) between two projection vectors. For LDA algorithms, the distance can be unweighted or weighted.

In Swets and Weng [1996b], discriminant analysis of eigenfeatures is applied in an image retrieval system to determine not only class (human face vs. nonface objects) but also individuals within the face class. Using tree-structure learning, the eigenspace and LDA projections are recursively applied to smaller and smaller sets of samples. Such recursive partitioning is carried out for every node until the samples assigned to the node belong to a single class. Experiments on this approach were reported in Swets and Weng [1996]. A set of 800 images was used for training; the training set came from 42 classes, of which human faces belong to a single class. Within the single face class, 356 individuals were included and distinguished. Testing results on images not in the training set were 91% for 78 face images and 87% for 38 nonface images based on the top choice.

A comparative performance analysis was carried out in Belhumeur et al. [1997]. Four methods were compared in this paper: (1) a correlation-based method, (2) a variant of the linear subspace method suggested in Shashua [1994], (3) an eigenface method Turk and Pentland [1991], and (4) a Fisherface method which uses subspace projection prior to LDA projection to avoid the possible singularity in S_w as in Swets and Weng [1996b]. Experiments were performed on a database of 500 images created by Hallinan [1994] and a database of 176 images created at Yale. The results of the experiments showed that the Fisherface method performed significantly better than the other three methods. However, no claim was made about the relative performance of these algorithms on larger databases.

To improve the performance of LDA-based systems, a regularized subspace LDA system that unifies PCA and LDA was proposed in Zhao [1999] and Zhao et al. [1998]. Good generalization ability of this system was demonstrated by experiments that carried out testing on new classes/individuals without retraining the PCA bases Φ, and sometimes the LDA bases W. While the reason for not retraining PCA is obvious, it is interesting to test the adaptive capability of the system by fixing the LDA bases when images from new classes are added.[7] The fixed PCA subspace of dimensionality 300 was trained from a large number of samples. An augmented set of 4056 mostly frontal-view images constructed from the original 1078 FERET images of 444 individuals by adding noise and mirroring was used in Zhao et al. [1998]. At least one of the following three characteristics separates this system from other LDA-based systems: (1) the unique selection of the universal face subspace dimension, (2) the use of a weighted distance measure, and (3) a regularized procedure that modifies the within-class scatter matrix S_w. The authors selected the dimensionality of the universal face subspace based on the characteristics of the eigenvectors (face-like or not) instead of the eigenvalues [Zhao et al. 1998], as is commonly done. Later it was concluded in Penev and Sirovich [2000] that the global face subspace dimensionality is on the order of 400 for large databases of 5,000 images. A weighted distance metric in the projection space z was used to improve performance [Zhao 1999].[8] Finally, the LDA

[7]This makes sense because the final classification is carried out in the projection space z by comparison with prestored projection vectors.

[8]Weighted metrics have also been used in the pure LDA approach [Etemad and Chellappa 1997] and the

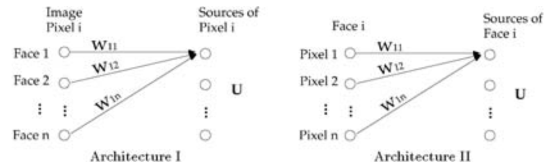

Fig. 7. Two architectures for performing ICA on images. Left: architecture for finding statistically independent basis images. Performing source separation on the face images produces independent images in the rows of U. Right: architecture for finding a factorial code. Performing source separation on the pixels produces a factorial code in the columns of the output matrix U [Bartlett et al. 1998]. (Courtesy of M. Bartlett, H. Lades, and T. Sejnowski.)

training was regularized by modifying the S_w matrix to $S_w+\delta I$, where δ is a relatively small positive number. Doing this solves a numerical problem when S_w is close to being singular. In the extreme case where only one sample per class is available, this regularization transforms the LDA problem into a standard PCA problem with S_b being the covariance matrix C. Applying this approach, without retraining the LDA basis, to a testing/probe set of 46 individuals of which 24 were trained and 22 were not trained (a total of 115 images including 19 untrained images of nonfrontal views), the authors reported the following performance based on a front-view-only gallery database of 738 images: 85.2% for all images and 95.1% for frontal views.

An evolution pursuit- (EP-) based adaptive representation and its application to face recognition were presented in Liu and Wechsler [2000a]. In analogy to projection pursuit methods, EP seeks to learn an optimal basis for the dual purpose of data compression and pattern classification. In order to increase the generalization ability of EP, a balance is sought between minimizing the empirical risk encountered during training and narrowing the confidence interval for reducing the guaranteed risk during future testing on unseen data [Vapnik 1995]. Toward that end, EP implements strategies characteristic of genetic algorithms (GAs) for searching the space of possible solutions to determine the optimal basis. EP starts by projecting the original data into a lower-dimensional whitened PCA space. Directed random rotations of the basis vectors in this space are then searched by GAs where evolution is driven by a fitness function defined in terms of performance accuracy (empirical risk) and class separation (confidence interval). The feasibility of this method has been demonstrated for face recognition, where the large number of possible bases requires a greedy search algorithm. The particular face recognition task involves 1107 FERET frontal face images of 369 subjects; there were three frontal images for each subject, two for training and the remaining one for testing. The authors reported improved face recognition performance as compared to eigenfaces [Turk and Pentland 1991], and better generalization capability than Fisherfaces [Belhumeur et al. 1997].

Based on the argument that for tasks such as face recognition much of the important information is contained in high-order statistics, it has been proposed [Bartlett et al. 1998] to use ICA to extract features for face recognition. Independent-component analysis is a generalization of principal-component analysis, which decorrelates the high-order moments of the input in addition to the second-order moments. Two architectures have been proposed for face recognition (Figure 7): the first is used to find a set of statistically independent source images

so-called enhanced FLD (EFM) approach [Liu and Wechsler 2000b].

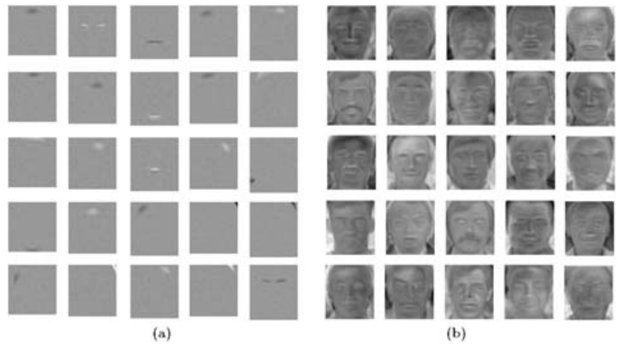

Fig. 8. Comparison of basis images using two architectures for performing ICA: (a) 25 independent components of Architecture I, (b) 25 independent components of Architecture II [Bartlett et al. 1998]. (Courtesy of M. Bartlett, H. Lades, and T. Sejnowski.)

that can be viewed as independent image features for a given set of training images [Bell and Sejnowski 1995], and the second is used to find image filters that produce statistically independent outputs (a factorial code method) [Bell and Sejnowski 1997]. In both architectures, PCA is used first to reduce the dimensionality of the original image size (60×50). ICA is performed on the first 200 eigenvectors in the first architecture, and is carried out on the first 200 PCA projection coefficients in the second architecture. The authors reported performance improvement of both architectures over eigenfaces in the following scenario: a FERET subset consisting of 425 individuals was used; all the frontal views (one per class) were used for training and the remaining (up to three) frontal views for testing. Basis images of the two architectures are shown in Figure 8 along with the corresponding eigenfaces.

3.2.1.2. Other Representations. In addition to the popular PCA representation and its derivatives such as ICA and EP, other features have also been used, such as raw intensities and edges.

A fully automatic face detection/recognition system based on a neural network is reported in Lin et al. [1997]. The proposed system is based on a probabilistic decision-based neural network (PDBNN, an extended (DBNN) [Kung and Taur 1995]) which consists of three modules: a face detector, an eye localizer, and a face recognizer. Unlike most methods, the facial regions contain the eyebrows, eyes, and nose, but not the mouth.[9] The rationale of using only the upper face is to build a robust system that excludes the influence of facial variations due to expressions that cause motion around the mouth. To improve robustness, the segmented facial region images are first processed to produce two features at a reduced resolution of 14×10: normalized intensity features and edge features, both in the range $[0, 1]$. These features are fed into two PDBNNs and the final recognition result is the fusion of the outputs of these two PDBNNs. A unique characteristic of PDBNNs and DBNNs is their modular structure. That is, for each class/person

[9]Such a representation was also used in Kirby and Sirovich [1990]

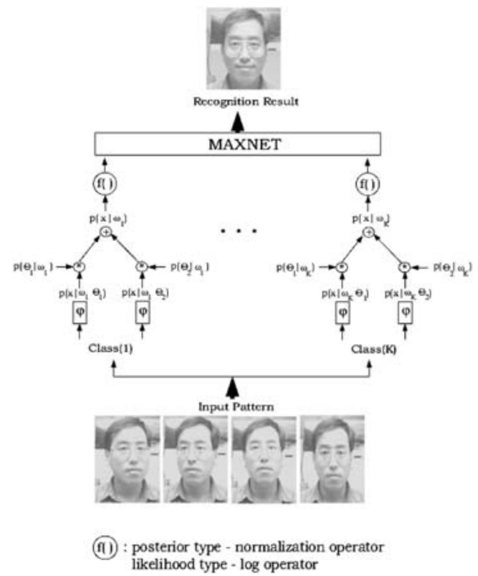

Fig. 9. Structure of the PDBNN face recognizer. Each class subnet is designed to recognize one person. All the network weightings are in probabilistic format [Lin et al. 1997]. (Courtesy of S. Lin, S. Kung, and L. Lin.)

to be recognized, PDBNN/DBNN devotes one of its subnets to the representation of that particular person, as illustrated in Figure 9. Such a one-class-in-one-network (OCON) structure has certain advantages over the all-classes-in-one-network (ACON) structure that is adopted by the conventional multilayer perceptron (MLP). In the ACON structure, all classes are lumped into one supernetwork, so large numbers of hidden units are needed and convergence is slow. On the other hand, the OCON structure consists of subnets that consist of small numbers of hidden units; hence it not only converges faster but also has better generalization capability. Compared to most multiclass recognition systems that use a discrimination function between any two classes, PDBNN has a lower false acceptance/rejection rate because it uses the full density description for each class. In addition, this architecture is beneficial for hardware implementation such as distributed computing. However, it is not clear how to accurately estimate the full density functions for the classes when there are only limited numbers of samples. Further, the system could have problems when the number of classes grows exponentially.

3.2.2. Feature-Based Structural Matching Approaches. Many methods in the structural matching category have been proposed, including many early methods based on geometry of local features [Kanade 1973;

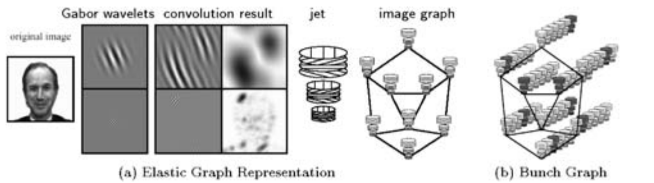

Fig. 10. The bunch graph representation of faces used in elastic graph matching [Wiskott et al. 1997]. (Courtesy of L. Wiskott, J.-M. Fellous, and C. von der Malsburg.)

Kelly 1970] as well as 1D [Samaria and Young 1994] and pseudo-2D [Samaria 1994] HMM methods. One of the most successful of these systems is the Elastic Bunch Graph Matching (EBGM) system [Okada et al. 1998; Wiskott et al. 1997], which is based on DLA [Buhmann et al. 1990; Lades et al. 1993]. Wavelets, especially Gabor wavelets, play a building block role for facial representation in these graph matching methods. A typical local feature representation consists of wavelet coefficients for different scales and rotations based on fixed wavelet bases (called *jets* in Okada et al. [1998]). These locally estimated wavelet coefficients are robust to illumination change, translation, distortion, rotation, and scaling.

The basic 2D Gabor function and its Fourier transform are

$$
\begin{aligned}
g(x, y : u_0, v_0) &= \exp\left(-[x^2/2\sigma_x^2 + y^2/2\sigma_y^2] + 2\pi i[u_0 x + v_o y]\right),\\
G(u, v) &= \exp\left(-2\pi^2(\sigma_x^2(u-u_0)^2 + \sigma_y^2(v-v_0)^2)\right), \qquad (5)
\end{aligned}
$$

where σ_x and σ_y represent the spatial widths of the Gaussian and (u_0, v_0) is the frequency of the complex sinusoid.

DLAs attempt to solve some of the conceptual problems of conventional artificial neural networks, the most prominent of these being the representation of syntactical relationships in neural networks. DLAs use synaptic plasticity and are able to form sets of neurons grouped into structured graphs while maintaining the advantages of neural systems. Both Buhmann et al. [1990] and Lades et al. [1993] used Gabor-based wavelets (Figure 10(a)) as the features. As described in Lades et al. [1993] DLA's basic mechanism, in addition to the connection parameter T_{ij} betweeen two neurons (i, j), is a dynamic variable J_{ij}. Only the J-variables play the roles of synaptic weights for signal transmission. The T-parameters merely act to constrain the J-variables, for example, $0 \le J_{ij} \le T_{ij}$. The T-parameters can be changed slowly by long-term synaptic plasticity. The weights J_{ij} are subject to rapid modification and are controlled by the signal correlations between neurons i and j. Negative signal correlations lead to a decrease and positive signal correlations lead to an increase in J_{ij}. In the absence of any correlation, J_{ij} slowly returns to a resting state, a fixed fraction of T_{ij}. Each stored image is formed by picking a rectangular grid of points as graph nodes. The grid is appropriately positioned over the image and is stored with each grid point's locally determined jet (Figure 10(a)), and serves to represent the pattern classes. Recognition of a new image takes place by transforming the image into the grid of jets, and matching all stored model graphs to the image. Conformation of the DLA is done by establishing and dynamically modifying links between vertices in the model domain.

The DLA architecture was recently extended to Elastic Bunch Graph Matching [Wiskott et al. 1997] (Figure 10). This is similar to the graph described above, but instead of attaching only a single jet to each node, the authors attached a set of jets (called the *bunch graph representation*, Figure 10(b)), each derived from a different face image. To handle the pose variation problem, the pose of the face is first determined using prior class information [Kruger et al. 1997], and the "jet" transformations under pose variation are learned [Maurer and Malsburg 1996a]. Systems based on the EBGM approach have been applied to face detection and extraction, pose estimation, gender classification, sketch-image-based recognition, and general object recognition. The success of the EBGM system may be due to its resemblance to the human visual system [Biederman and Kalocsai 1998].

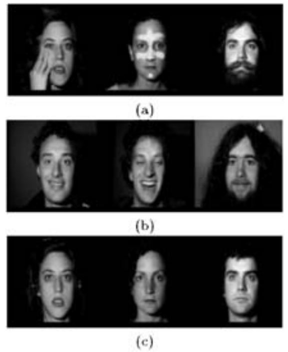

Fig. 11. Comparison of matching: (a) test views, (b) eigenface matches, (c) eigenfeature matches [Pentland et al. 1994].

3.2.3. Hybrid Approaches. Hybrid approaches use both holistic and local features. For example, the modular eigenfaces approach [Pentland et al. 1994] uses both global eigenfaces and local eigenfeatures.

In Pentland et al. [1994], the capabilities of the earlier system [Turk and Pentland 1991] were extended in several directions. In mugshot applications, usually a frontal and a side view of a person are available; in some other applications, more than two views may be appropriate. One can take two approaches to handling images from multiple views. The first approach pools all the images and constructs a set of eigenfaces that represent all the images from all the views. The other approach uses separate eigenspaces for different views, so that the collection of images taken from each view has its own eigenspace. The second approach, known as *view-based eigenspaces*, performs better.

The concept of eigenfaces can be extended to eigenfeatures, such as eigeneyes, eigenmouth, etc. Using a limited set of images (45 persons, two views per person, with different facial expressions such as neutral vs. smiling), recognition performance as a function of the number of eigenvectors was measured for eigenfaces only and for the combined representation. For lower-order spaces, the eigenfeatures performed better than the eigenfaces [Pentland et al. 1994]; when the combined set was used, only marginal improvement was obtained. These experiments support the claim that feature-based mechanisms may be useful when gross variations are present in the input images (Figure 11).

It has been argued that practical systems should use a hybrid of PCA and LFA (Appendix B in Penev and Atick [1996]). Such view has been long held in the psychology community [Bruce 1988]. It seems to be better to estimate eigenmodes/eigenfaces that have large eigenvalues (and so are more robust against noise), while for estimating higher-order eigenmodes it is better to use LFA. To support this point, it was argued in Penev and Atick [1996] that the leading eigenpictures are global, integrating, or smoothing filters that are efficient in suppressing noise, while the higher-order modes are ripply or differentiating filters that are likely to amplify noise.

LFA is an interesting biologically inspired feature analysis method [Penev and Atick 1996]. Its biological motivation comes from the fact that, though a huge array of receptors (more than six million cones) exist in the human retina, only a

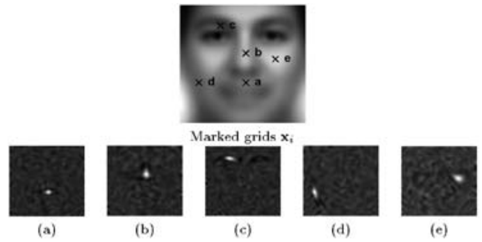

Fig. 12. LFA kernels $K(\mathbf{x}_i, \mathbf{y})$ at different grids $\mathbf{x}_i$ [Penev and Atick 1996].

small fraction of them are active, corresponding to natural objects/signals that are statistically redundant [Ruderman 1994]. From the activity of these sparsely distributed receptors, the brain has to discover where and what objects are in the field of view and recover their attributes. Consequently, one expects to represent the natural objects/signals in a subspace of lower dimensionality by finding a suitable parameterization. For a limited class of objects such as faces which are correctly aligned and scaled, this suggests that even lower dimensionality can be expected [Penev and Atick 1996]. One good example is the successful use of the truncated PCA expansion to approximate the frontal face images in a linear subspace [Kirby and Sirovich 1990; Sirovich and Kirby 1987].

Going a step further, the whole face region stimulates a full 2D array of receptors, each of which corresponds to a location in the face, but some of these receptors may be inactive. To explore this redundancy, LFA is used to extract topographic local features from the global PCA modes. Unlike PCA kernels Φ_i which contain no topographic information (their supports extend over the entire grid of images), LFA kernels (Figure12) $K(\mathbf{x}_i, \mathbf{y})$ at selected grids $\mathbf{x}_i$ have local support.[10]

[10]These kernels (Figure 12) indexed by grids $\mathbf{x}_i$ are similar to the ICA kernels in the first ICA system architecture [Bartlett et al. 1998; Bell and Sejnowski 1995].

The search for the best topographic set of sparsely distributed grids $\{x_o\}$ based on reconstruction error is called *sparsification* and is described in Penev and Atick [1996]. Two interesting points are demonstrated in this paper: (1) using the same number of kernels, the perceptual reconstruction quality of LFA based on the optimal set of grids is better than that of PCA; the mean square error is 227, and 184 for a particular input; (2) keeping the second PCA eigenmodel in LFA reconstruction reduces the mean square error to 152, suggesting the hybrid use of PCA and LFA. No results on recognition performance based on LFA were reported. LFA is claimed to be used in Visionics's commercial system FaceIt (Table II).

A flexible appearance model based method for automatic face recognition was presented in [Lanitis et al. 1995]. To identify a face, both shape and gray-level information are modeled and used. The shape model is an ASM; these are statistical models of the shapes of objects which iteratively deform to fit to an example of the shape in a new image. The statistical shape model is trained on example images using PCA, where the variables are the coordinates of the shape model points. For the purpose of classification, the shape variations due to interclass variation are separated from those due to within-class variations (such as small variations in 3D orientation and facial expression) using discriminant analysis. Based on the average shape of the

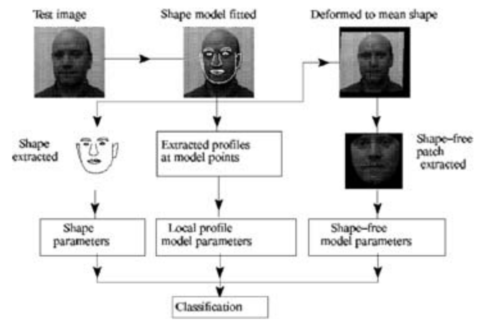

Fig. 13. The face recognition scheme based on flexible appearance model [Lanitis et al. 1995]. (Courtesy of A. Lanitis, C. Taylor, and T. Cootes.)

shape model, a global shape-free gray-level model can be constructed, again using PCA.[11] To further enhance the robustness of the system against changes in local appearance such as occlusions, local gray-level models are also built on the shape model points. Simple local profiles perpendicular to the shape boundary are used. Finally, for an input image, all three types of information, including extracted shape parameters, shape-free image parameters, and local profiles, are used to compute a Mahalanobis distance for classification as illustrated in Figure 13. Based on training 10 and testing 13 images for each of 30 individuals, the classification rate was 92% for the 10 normal testing images and 48% for the three difficult images.

The last method [Huang et al. 2003] that we review in this category is based on recent advances in component-based detection/recognition [Heisele et al. 2001] and 3D morphable models [Blanz and Vetter 1999]. The basic idea of component-based methods [Heisele et al. 2001] is to decompose a face into a set of facial components such as mouth and eyes that are interconnected by a flexible geometrical model. (Notice how this method is similar to the EBGM system [Okada et al. 1998; Wiskott et al. 1997] except that gray-scale components are used instead of Gabor wavelets.) The motivation for using components is that changes in head pose mainly lead to changes in the positions of facial components which could be accounted for by the flexibility of the geometric model. However, a major drawback of the system is that it needs a large number of training images taken from different viewpoints and under different lighting conditions. To overcome this problem, the 3D morphable face model [Blanz and Vetter 1999] is applied to generate arbitrary synthetic images under varying pose and illumination. Only three face images (frontal, semiprofile, profile) of a person are needed to compute the 3D face model. Once the 3D model is constructed, synthetic images of size 58×58 are generated for training both the detector and the classifer. Specifically, the faces were rotated in depth from 0° to 34° in 2° increments and rendered with two illumination models (the first model consists of ambient light alone and the second includes ambient light and a rotating point light source) at each pose. Fourteen facial components were used for face detection, but only nine components

[11]Recall that in Craw and Cameron [1996] and Moghaddam and Pentland [1997] these shape-free images are used as the inputs to the classifier.

that were not strongly overlapped and contained gray-scale structures were used for classification. In addition, the face region was added to the nine components to form a single feature vector (a hybrid method), which was later trained by a SVM classifer [Vapnik 1995]. Training on three images and testing on 200 images per subject led to the following recognition rates on a set of six subjects: 90% for the hybrid method and roughly 10% for the global method that used the face region only; the false positive rate was 10%.

3.3. Summary and Discussion

Face recognition based on still images or captured frames in a video stream can be viewed as 2D image matching and recognition; range images are not available in most commercial/law enforcement applications. Face recognition based on other sensing modalities such as sketches and infrared images is also possible. Even though this is an oversimplification of the actual recognition problem of 3D objects based on 2D images, we have focused on this 2D problem, and we will address two important issues about 2D recognition of 3D face objects in Section 6. Significant progress has been achieved on various aspects of face recognition: segmentation, feature extraction, and recognition of faces in intensity images. Recently, progress has also been made on constructing fully automatic systems that integrate all these techniques.

3.3.1. Status of Face Recognition. After more than 30 years of research and development, basic 2D face recognition has reached a mature level and many commercial systems are available (Table II) for various applications (Table I).

Early research on face recognition was primarily focused on the feasibility question, that is: is machine recognition of faces possible? Experiments were usually carried out using datasets consisting of as few as 10 images. Significant advances were made during the mid-1990s, with many methods proposed and tested on datasets consisting of as many as 100 images. More recently, practical methods have emerged that aim at more realistic applications. In the recent comprehensive FERET evaluations [Phillips et al. 2000; Phillips et al. 1998b; Rizvi et al. 1998], aimed at evaluating different systems using the same large database containing thousands of images, the systems described in Moghaddam and Pentland [1997]; Swets and Weng [1996b]; Turk and Pentland [1991]; Wiskott et al. [1997]; Zhao et al. [1998], as well as others, were evaluated. The EBGM system [Wiskott et al. 1997], the subspace LDA system [Zhao et al. 1998], and the probabilistic eigenface system [Moghaddam and Pentland 1997] were judged to be among the top three, with each method showing different levels of performance on different subsets of sequestered images. A brief summary of the FERET evaluations will be presented in Section 5. Recently, more extensive evaluations using commercial systems and thousands of images have been performed in the FRVT 2000 [Blackburn et al. 2001] and FRVT 2002 [Phillips et al. 2003] tests.

3.3.2. Lessons, Facts and Highlights. During the development of face recognition systems, many lessons have been learned which may provide some guidance in the development of new methods and systems.

—Advances in face recognition have come from considering various aspects of this specialized perception problem. Earlier methods treated face recognition as a standard pattern recognition problem; later methods focused more on the representation aspect, after realizing its uniqueness (using domain knowledge); more recent methods have been concerned with both representation and recognition, so a robust system with good generalization capability can be built. Face recognition continues to adopt state-of-the-art techniques from learning, computer vision, and pattern recognition. For example, distribution modeling using mixtures of Gaussians, and SVM learning methods, have been used in face detection/recognition.

—Among all face detection/recognition methods, appearance/image-based approaches seem to have dominated up to now. The main reason is the strong prior that all face images belong to a face class. An important example is the use of PCA for the representation of holistic features. To overcome sensitivity to geometric change, local appearance-based approaches, 3D enhanced approaches, and hybrid approaches can be used. The most recent advances toward fast 3D data acquisition and accurate 3D recognition are likely to influence future developements.[12]

—The methodological difference between face detection and face recognition may not be as great as it appears to be. We have observed that the multiclass face recognition problem can be converted into a two-class "detection" problem by using image differences [Moghaddam and Pentland 1997]; and the face detection problem can be converted into a multiclass "recognition" problem by using additional nonface clusters of negative samples [Sung and Poggio 1997].

—It is well known that for face detection, the image size can be quite small. But what about face recognition? Clearly the image size cannot be too small for methods that depend heavily on accurate feature localization, such as graph matching methods [Okada et al. 1998]. However, it has been demonstrated that the image size can be very small for holistic face recognition: 12×11 for the subspace LDA system [Zhao et al. 1999], 14×10 for the PDBNN system [Lin et al. 1997], and 18×24 for human perception [Bachmann 1991]. Some authors have argued that there exists a universal face subspace of fixed dimension; hence for holistic recognition, image size does not matter as long as it exceeds the subspace dimensionality [Zhao et al. 1999]. This claim has been supported by limited experiments using normalized face images of different sizes, for example, from 12×11 to 48×42, to obtain different face subspaces [Zhao 1999]. Indeed, slightly better performance was observed when smaller images were used. One reason is that the signal-to-noise ratio improves with the decrease in image size.

—Accurate feature location is critical for good recognition performance. This is true even for holistic matching methods, since accurate location of key facial features such as eyes is required to normalize the detected face [Yang et al. 2002; Zhao 1999]. This was also verified in Lin et al. [1997] where the use of smaller images led to slightly better performance due to increased tolerance to location errors. In Martinez [2002], a systematic study of this issue was presented.

—Regarding the debate in the psychology community about whether face recognition is a dedicated process, the recent success of machine systems that are trained on large numbers of samples seems to confirm recent findings suggesting that human recognition of faces may be not unique/dedicated, but needs extensive training.

—When comparing different systems, we should pay close attention to implementation details. Different implementations of a PCA-based face recognition algorithm were compared in Moon and Phillips [2001]. One class of variations examined was the use of seven different distance metrics in the nearest-neighbor classifier, which was found to be the most critical element. This raises the question of what is more important in algorithm performance, the representation or the specifics of the implementation. Implementation details often determine the performance of a system. For example, input images are normalized only with respect to translation, in-plane rotation, and scale in Belhumeur et al. [1997], Swets and Weng [1996b], Turk and Pentland [1991], and Zhao et al. [1998], whereas in Moghaddam and Pentland [1997] the normalization also includes masking and affine warping to align the shape. In Craw and Cameron [1996], manually selected points are used to warp the input images to the mean shape, yielding shape-free images. Because of this difference, PCA was a good classifier in Moghaddam and Pentland [1997] for the shape-free representations, but it may not be as good for the simply normalized representations. Recently, systematic comparisons and independent reevaluations of existing methods have been published [Beveridge et al. 2001]. This is beneficial to the research community. However, since the methods need to be reimplemented, and not all the details in the original implementation can be taken into account, it is difficult to carry out absolutely fair comparisons.

—Over 30 years of research has provided us with a vast number of methods and systems. Recognizing the fact that each method has its advantages and disadvantages, we should select methods and systems appropriate to the application. For example, local feature based methods cannot be applied when the input image contains a small face region, say 15×15. Another issue is when to use PCA and when to use LDA in building a system. Apparently, when the number of training samples per class is large, LDA is the best choice. On the other hand, if only one or two samples are available per class (a degenerate case for LDA), PCA is a better choice. For a more detailed comparison of PCA versus LDA, see Beveridge et al. [2001]; Martinez and Kak [2001]. One way to unify PCA and LDA is to use regularized subspace LDA [Zhao et al. 1999].

[12]Early work using range images was reported in Gordon [1991].

3.3.3. Open Research Issues. Though machine recognition of faces from still images has achieved a certain level of success, its performance is still far from that of human perception. Specifically, we can list the following open issues:

—Hybrid face recognition systems that use both holistic and local features resemble the human perceptual system. While the holistic approach provides a quick recognition method, the discriminant information that it provides may not be rich enough to handle very large databases. This insufficiency can be compensated for by local feature methods. However, many questions need to be answered before we can build such a combined system. One important question is how to arbitrate the use of holistic and local features. As a first step, a simple, naive engineering approach would be to weight the features. But how to determine *whether* and *how to* use the features remains an open problem.

—The challenge of developing face detection techniques that report not only the presence of a face but also the accurate locations of facial features under large pose and illumination variations still remains. Without accurate localization of important features, accurate and robust face recognition cannot be achieved.

—How to model face variation under realistic settings is still challenging—for example, outdoor environments, natural aging, etc.

4. FACE RECOGNITION FROM IMAGE SEQUENCES

A typical video-based face recognition system automatically detects face regions, extracts features from the video, and recognizes facial identity if a face is present. In surveillance, information security, and access control applications, face recognition and identification from a video sequence is an important problem. Face recognition based on video is preferable over using still images, since as demonstrated in Bruce et al. [1998] and Knight and Johnston [1997], motion helps in recognition of (familiar) faces when the images are negated, inverted or threshold. It was also demonstrated that humans can recognize animated faces better than randomly rearranged images from the same set. Though recognition of faces from video sequence is a direct extension of still-image-based recognition, in our opinion, *true* video-based face recognition techniques that coherently use both spatial and temporal information started only a few years ago and still need further investigation. Significant challenges for video-based recognition still exist; we list several of them here.

(1) *The quality of video is low*. Usually, video acquisition occurs outdoors (or indoors but with bad conditions for video capture) and the subjects are not cooperative; hence there may be large illumination and pose variations in the face images. In addition, partial occlusion and disguise are possible.
(2) *Face images are small*. Again, due to the acquisition conditions, the face image sizes are smaller (sometimes much smaller) than the assumed sizes in most still-image-based face recognition systems. For example, the valid face region can be as small as 15×15 pixels,[13] whereas the face image sizes used in feature-based still image-based systems can be as large as 128×128. Small-size images not only make the recognition task more difficult, but also affect the accuracy of face segmentation, as well as the accurate detection of the fiducial points/landmarks that are often needed in recognition methods.
(3) *The characteristics of faces/human body parts*. During the past 8 years, research on human action/behavior recognition from video has been very active and fruitful. Generic description of human behavior not particular to an individual is an interesting and useful concept. One of the main reasons for the feasibility of generic descriptions of human behavior is that the intraclass variations of human bodies, and in particular faces, is much smaller than the difference between the objects inside and outside the class. For the same reason, recognition of individuals within the class is difficult. For example, detecting and localizing faces is typically much easier than recognizing a specific face.

[13]Notice this is totally different from the situation where we have images with large face regions but the final face regions feed into a classifier is 15×15.

Before we examine existing video-based face recognition algorithms, we briefly review three closely related techniques: face segmentation and pose estimation, face tracking, and face modeling. These techniques are critical for the realization of the full potential of video-based face recognition.

4.1. Basic Techniques of Video-Based Face Recognition

In Chellappa et al. [1995], four computer vision areas were mentioned as being important for video-based face recognition: segmentation of moving objects (humans) from a video sequence; structure estimation; 3D models for faces; and nonrigid motion analysis. For example, in Jebara et al. [1998] a face modeling system which estimates facial features and texture from a video stream was described. This system utilizes all four techniques: segmentation of the face based on skin color to initiate tracking; use of a 3D face model based on laser-scanned range data to normalize the image (by facial feature alignment and texture mapping to generate a frontal view) and construction of an eigen-subspace for 3D heads; use of structure from motion (SfM) at each feature point to provide depth information; and nonrigid motion analysis of the facial features based on simple 2D SSD (sum of squared differences) tracking constrained by a global 3D model. Based on the current development of video-based face recognition, we think it is better to review three specific face-related techniques instead of the above four general areas. The three video-based face-related techniques are: face segmentation and pose estimation, face tracking, and face modeling.

4.1.1. Face Segmentation and Pose Estimation. Early attempts [Turk and Pentland 1991] at segmenting moving faces from an image sequence used simple pixel-based change detection procedures based on difference images. These techniques may run into difficulties when multiple moving objects and occlusion are present. More sophisticated methods use estimated flow

fields for segmenting humans in motion [Shio and Sklansky 1991]. More recent methods [Choudhury et al. 1999; McKenna and Gong 1998] have used motion and/or color information to speed up the process of searching for possible face regions. After candidate face regions are located, still-image-based face detection techniques can be applied to locate the faces [Yang et al. 2002]. Given a face region, important facial features can be located. The locations of feature points can be used for pose estimation, which is important for synthesizing a virtual frontal view [Choudhury et al. 1999]. Newly developed segmentation methods locate the face and estimate its pose simultaneously without extracting features [Gu et al. 2001; Li et al. 2001b]. This is achieved by learning multiview face examples which are labeled with manually determined pose angles.

4.1.2. Face and Feature Tracking. After faces are located, the faces and their features can be tracked. Face tracking and feature tracking are critical for reconstructing a face model (depth) through SfM, and feature tracking is essential for facial expression recognition and gaze recognition. Tracking also plays a key role in spatiotemporal-based recognition methods [Li and Chellappa 2001; Li et al. 2001a] which directly use the tracking information.

In its most general form, tracking is essentially motion estimation. However, general motion estimation has fundamental limitations such as the aperture problem. For images like faces, some regions are too smooth to estimate flow accurately, and sometimes the change in local appearances is too large to give reliable flow. Fortunately, these problems are alleviated thanks to face modeling, which exploits domain knowledge. In general, tracking and modeling are dual processes: tracking is constrained by a generic 3D model or a learned statistical model under deformation, and individual models are refined through tracking. Face tracking can be roughly divided into three categories: (1) head tracking, which involves tracking the motion of a rigid object that is performing rotations and translations; (2) facial feature tracking, which involves tracking nonrigid deformations that are limited by the anatomy of the head, that is, articulated motion due to speech or facial expressions and deformable motion due to muscle contractions and relaxations; and (3) complete tracking, which involves tracking both the head and the facial features.

Early efforts focused on the first two problems: head tracking [Azarbayejani et al. 1993] and facial feature tracking [Terzopoulos and Waters 1993; Yuille and Hallinan 1992]. In Azarbayejani et al. [1993], an approach to head tracking using points with high Hessian values was proposed. Several such points on the head are tracked and the 3D motion parameters of the head are recovered by solving an overconstrained set of motion equations. Facial feature tracking methods may make use of the feature boundary or the feature region. Feature boundary tracking attempts to track and accurately delineate the shape of the facial feature, for example, to track the contours of the lips and mouth [Terzopoulos and Waters 1993]. Feature region tracking addresses the simpler problem of tracking a region such as a bounding box that surrounds the facial feature [Black et al. 1995].

In Black et al. [1995], a tracking system based on local parameterized models is used to recognize facial expressions. The models include a planar model for the head, local affine models for the eyes, and local affine models and curvature for the mouth and eyebrows. A face tracking system was used in Maurer and Malsburg [1996b] to estimate the pose of the face. This system used a graph representation with about 20–40 nodes/landmarks to model the face. Knowledge about faces is used to find the landmarks in the first frame. Two tracking systems described in Jebara et al. [1998] and Strom et al. [1999] model faces completely with texture and geometry. Both systems use generic 3D models and SfM to recover the face structure. Jebara et al. [1998] relied fixed feature points (eyes, nose tip),

Table IV. Categorization of Video-Based Face Recognition Techniques

Approach	Representative work
Still-image methods	Basic methods [Turk and Pentland 1991; Lin et al. 1997; Moghaddam and Pentland 1997; Okada et al. 1998; Penev and Atick 1996; Wechsler et al. 1997; Wiskott et al. 1997]
	Tracking-enhanced [Edwards et al. 1998; McKenna and Gong 1997, 1998; Steffens et al. 1998]
Multimodal methods	Video- and audio-based [Bigun et al. 1998; Choudhury et al. 1999]
Spatiotemporal methods	Feature trajectory-based [Li and Chellappa 2001; Li et al. 2001a]
	Video-to video methods [Zhou et al. 2003]

while Strom et al. [1999] tracked only points with high Hessian values. Also, Jebara et al. [1998] tracked 2D features in 3D by deforming them, while Strom et al. [1999] relied on direct comparison of a 3D model to the image. Methods have been proposed in Black et al. [1998] and Hager and Belhumeur [1998] to solve the varying appearance (both geometry and photometry) problem in tracking. Some of the newest model-based tracking methods calculate the 3D motions and deformations directly from image intensities [Brand and Bhotika 2001], thus eliminating the information-lossy intermediate representations.

4.1.3. Face Modeling. Modeling of faces includes 3D shape modeling and texture modeling. For large texture variations due to changes in illumination, we will address the illumination problem in Section 6. Here we focus on 3D shape modeling. 3D models of faces have been employed in the graphics, animation, and model-based image compression literature. More complicated models are used in applications such as forensic face reconstruction from partial information.

In computer vision, one of the most widely used methods of estimating 3D shape from a video sequence is SfM, which estimates the 3D depths of interesting points. The unconstrained SfM problem has been approached in two ways. In the differential approach, one computes some type of flow field (optical, image, or normal) and uses it to estimate the depths of visible points. The difficulty in this approach is reliable computation of the flow field. In the discrete approach, a set of features such as points, edges, corners, lines, or contours are tracked over a sequence of frames, and the depths of these features are computed. To overcome the difficulty of feature tracking, bundle adjustment [Triggs et al. 2000] can be used to obtain better and more robust results.

Recently, multiview based 2D methods have gained popularity. In Li et al. [2001b], a model consisted of a sparse 3D shape model learned from 2D images labeled with pose and landmarks, a shape-and-pose-free texture model, and an affine geometrical model. An alternative approach is to use 3D models such as the deformable model of DeCarlo and Metaxas [2000] or the linear 3D object class model of Blanz and Vetter [1999]. (In Blanz and Vetter [1999] a morphable 3D face model consisting of shape and texture was directly matched to single/multiple input images; as a consequence, head orientation, illumination conditions, and other parameters could be free variables subject to optimization.) In Blanz and Vetter [1999], real-time 3D modeling and tracking of faces was described; a generic 3D head model was aligned to match frontal views of the face in a video sequence.

4.2. Video-Based Face Recognition

Historically, video face recognition originated from still-image-based techniques (Table IV). That is, the system automatically detects and segments the face from the video, and then applies still-image face recognition techniques. Many methods reviewed in Section 3 belong to this category: eigenfaces [Turk and Pentland 1991], probabilistic eigenfaces [Moghaddam and Pentland 1997], the EBGM method [Okada et al. 1998; Wiskott et al. 1997], and the PDBNN method [Lin et al. 1997]. An improvement over these methods is to apply tracking; this can help

in recognition, in that a virtual frontal view can be synthesized via pose and depth estimation from video. Due to the abundance of frames in a video, another way to improve the recognition rate is the use of "voting" based on the recognition results from each frame. The voting can be deterministic, but probabilistic voting is better in general [Gong et al. 2000; McKenna and Gong 1998]. One drawback of such voting schemes is the expense of computing the deterministic/probabilistic results for each frame.

The next phase of video-based face recognition will be the use of multimodal cues. Since humans routinely use multiple cues to recognize identities, it is expected that a multimodal system will do better than systems based on faces only. More importantly, using multimodal cues offers a comprehensive solution to the task of identification that might not be achievable by using face images alone. For example, in a totally noncooperative environment, such as a robbery, the face of the robber is typically covered, and the only way to perform faceless identification might be to analyze body motion characteristics [Klasen and Li 1998]. Excluding fingerprints, face and voice are the most frequently used cues for identification. They have been used in many multimodal systems [Bigun et al. 1998; Choudhury et al. 1999]. Since 1997, a dedicated conference focused on video- and audio-based person authentication has been held every other year.

More recently, a third phase of video face recognition has started. These methods [Li and Chellappa 2001; Li et al. 2001a] coherently exploit both spatial information (in each frame) and temporal information (such as the trajectories of facial features). A big difference between these methods and the probabilistic voting methods [McKenna and Gong 1998] is the use of representations in a joint temporal and spatial space for identification.

We first review systems that apply still-image-based recognition to selected frames, and then multimodal systems. Finally, we review systems that use spatial and temporal information simultaneously.

In Wechsler et al. [1997], a fully automatic person authentication system was described which included video break, face detection, and authentication modules. Video skimming was used to reduce the number of frames to be processed. The video break module, corresponding to key-frame detection based on object motion, consisted of two units. The first unit implemented a simple optical flow method; it was used when the image SNR level was low. When the SNR level was high, simple pair-wise frame differencing was used to detect the moving object. The face detection module consisted of three units: face localization using analysis of projections along the x- and y-axes; face region labeling using a decision tree learned from positive and negative examples taken from 12 images each consisting of 2759 windows of size 8×8; and face normalization based on the numbers of face region labels. The normalized face images were then used for authentication, using an RBF network. This system was tested on three image sequences; the first was taken indoors with one subject present, the second was taken outdoors with two subjects, and the third was taken outdoors with one subject under stormy conditions. Perfect results were reported on all three sequences, as verified against a database of 20 still face images.

An access control system based on person authentication was described in McKenna and Gong [1997]. The system combined two complementary visual cues: motion and facial appearance. In order to reliably detect significant motion, spatiotemporal zero crossings computed from six consecutive frames were used. These motions were grouped into moving objects using a clustering algorithm, and Kalman filters were employed to track the grouped objects. An appearance-based face detection scheme using RBF networks (similar to that discussed in Rowley et al. [1998]) was used to confirm the presence of a person. The face detection scheme was "bootstrapped" using motion and object detection to provide an approximate head region. Face tracking based on the RBF network was used to provide feedback to the motion clustering process to help deal

Fig. 14. Varying the most significant identity parameters (top) and manipulating residual variation without affecting identity (bottom) [Edwards et al. 1998].

with occlusions. Good tracking results were demonstrated. In McKenna and Gong [1998], this work was extended to person authentication using PCA or LDA. The authors argued that recognition based on selected frames is not adequate since important information is discarded. Instead, they proposed a probabilistic voting scheme; that is, face identification was carried out continuously. Though they gave examples demonstrating improved performance in identifying 8 or 15 people by using sequences, no performance statistics were reported.

An appearance model based method for video tracking and enhancing identification was proposed in Edwards et al. [1998]. The appearance model is a combination of the active shape model (ASM) [Cootes et al. 1995] and the shape-free texture model after warping the face into a mean shape. Unlike Lanitis et al. [1995], which used the two models separately, the authors used a combined set of parameters for both models. The main contribution was the decomposition of the combined model parameters into an identity subspace and an orthogonal residual subspace using linear discriminant analysis. (See Figure 14 for an illustration of separating identity and residue.) The residual subspace would ideally contain intraperson variations caused by pose, lighting, and expression. In addition, they pointed out that optimal separation of identity and residue is class-specific. For example, the appearance change of a person's nose depends on its length, which is a person-specific quantity. To correct this class-specific information, a sequence of images of the same class was used. Specifically, a linear mapping was assumed to capture the relation between the class-specific correction to the identity subspace and the intraperson variation in the residual subspace. Examples of face tracking and visual enhancement were demonstrated, but no recognition experiments were reported. Though this method is believed to enhance tracking and make it robust against appearance change, it is not clear how efficient it is to learn the class-specific information from a video sequence that does not present much residual variation.

In De Carlo and Metaxas [2000], a system called PersonSpotter was described. This system is able to capture, track, and recognize a person walking toward or passing a stereo CCD camera. It has several modules, including a head tracker, preselector, landmark finder, and identifier. The head tracker determines the image regions that are changing due to object motion based on simple image differences. A stereo algorithm then determines the stereo disparities of these moving pixels. The disparity values are used to compute histograms for image regions. Regions within a certain disparity interval are selected and referred to as *silhouettes*. Two types of detectors, skin color based and convex region based, are applied to these silhouette images. The outputs of these detectors are clustered to form regions of interest which usually correspond to heads. To track a head robustly, temporal continuity is exploited in the form of

the thresholds used to initiate, track, and delete an object.

To find the face region in an image, the preselector uses a generic sparse graph consisting of 16 nodes learned from eight example face images. The landmark finder uses a dense graph consisting of 48 nodes learned from 25 example images to find landmarks such as the eyes and the nose tip. Finally, an elastic graph matching scheme is employed to identify the face. A recognition rate of about 90% was achieved; the size of the database is not known.

A multimodal person recognition system was described in Choudhury et al. [1999]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. It has the following characteristics: (1) the face recognition module can detect and compensate for pose variations; the speaker identification module can detect and compensate for changes in the auditory background; (2) the most reliable video frames and audio clips are selected for recognition; (3) 3D information about the head obtained through SfM is used to detect the presence of an actual person as opposed to an image of that person.

Two key parts of the face recognition module are face detection/tracking and eigen-face recognition. The face is detected using skin color information using a learned model of a mixture of Gaussians. The facial features are then located using symmetry transforms and image intensity gradients. Correlation-based methods are used to track the feature points. The locations of these feature points are used to estimate the pose of the face. This pose estimate and a 3D head model are used to warp the detected face image into a frontal view. For recognition, the feature locations are refined and the face is normalized with eyes and mouth in fixed locations. Images from the face tracker are used to train a frontal eigenspace, and the leading 35 eigenvectors are retained. Face recognition is then performed using a probabilistic eigenface approach where the projection coefficients of all images of each person are modeled as a Gaussian distribution.

Finally, the face and speaker recognition modules are combined using a Bayes net. The system was tested in an ATM scenario, a controlled environment. An ATM session begins when the subject enters the camera's field of view and the system detects his/her face. The system then greets the user and begins the banking transaction, which involves a series of questions by the system and answers by the user. Data for 26 people were collected; the normalized face images were 40×80 pixels and the audio was sampled at 16 kHz. These experiments on small databases and well-controlled environments showed that the combination of audio and video improved performance, and that 100% recognition and verification were achieved when the image/audio clips with highest confidence scores were used.

In Li and Chellappa [2001], a face verification system based on tracking facial features was presented. The basic idea of this approach is to exploit the temporal information available in a video sequence to improve face recognition. First, the feature points defined by Gabor attributes on a regular 2D grid are tracked. Then, the trajectories of these tracked feature points are exploited to identify the person presented in a short video sequence. The proposed tracking-for-verification scheme is different from the pure tracking scheme in that one template face from a database of known persons is selected for tracking. For each template with a specific personal ID, tracking can be performed and trajectories can be obtained. Based on the characteristics of these trajectories, identification can be carried out. According to the authors, the trajectories of the same person are more coherent than those of different persons, as illustrated in Figure 15. Such characteristics can also be observed in the posterior probabilities over time by assuming different classes. In other words, the posterior probabilities for the true hypothesis tend to be higher than those for false hypotheses. This in turn can be used for identification. Testing results on a small databases of 19

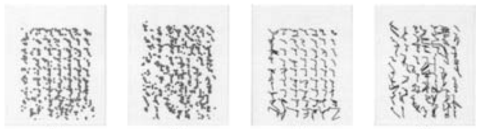

Fig. 15. Corresponding feature points obtained from 20 frames: (a) result of matching the same person to a video, (b) result of matching a different person to the video, (c) trajectories of (a), (d) trajectories of (b) [Li and Chellappa 2001].

individuals have suggested that performance is favorable over a frame-to-frame matching and voting scheme, especially in the case of large lighting changes. The testing result is based on comparison with alternative hypotheses.

Some details about the tracking algorithm are as follows [Li and Chellappa 2001]. The motion of facial feature points is modeled as a global two-dimensional (2D) affine transformation (accounting for head motion) plus a local deformation (accounting for residual motion that is due to inaccuracies in the 2D affine modeling and other factors such as facial expression). The tracking problem has been formulated as a Bayesian inference problem and sequential importance sampling (SIS) [Liu and Chen 1998] (one form of SIS is called *Condensation* [Isard and Blake 1996] in the computer vision literature) proposed as an empirical solution to the inference problem. Since SIS has difficulty in high-dimensional spaces, a reparameterization that captures essentially only the difference was used to facilitate the computation.

While most face recognition algorithms take still images as probe inputs, a video-based face recognition approach that takes video sequences as inputs has recently been developed [Zhou et al. 2003]. Since the detected face might be moving in the video sequence, one has to deal with uncertainty in tracking as well as in recognition. Rather than resolving these two uncertainties separately, Zhou et al. [2003] performed simultaneous tracking and recognition of human faces from a video sequence.

In still-to-video face recognition, where the gallery consists of still images, a time series state space model is proposed to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a motion vector and an identity variable, respectively. The joint posterior distribution of the motion vector and the identity variable is first estimated at each time instant and then propagated to the next time instant. Marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable and marginalization over the identity variable yields a robust estimate of the posterior distribution of the motion vector, so that tracking and recognition are handled simultaneously. A computationally efficient sequential importance sampling (SIS) algorithm is used to estimate the posterior distribution. Empirical results demonstrate that, due to the propagation of the identity variable over time, *degeneracy* in the posterior probability of the identity variable is achieved to give improved recognition. The gallery is generalized to videos in order to realize video-to-video face recognition. An exemplar-based learning strategy is employed to automatically select video representatives from the gallery, serving as mixture centers in an updated likelihood measure. The SIS algorithm is used to approximate the posterior distribution of the motion vector, the identity variable, and the exemplar index. The marginal distribution of the identity variable produces the recognition result. The model formulation is very general and allows a

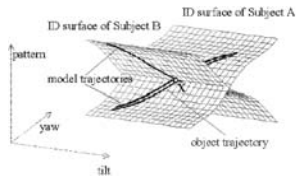

Fig. 16. Identity surface [Li et al. 2001a]. (Courtesy of Y. Li, S. Gong, and H. Liddell.)

variety of image representations and transformations. Experimental results using images/videos collected at UMD, NIST/USF, and CMU with pose/illumination variations have illustrated the effectiveness of this approach in both still-to-video and video-to-video scenarios with appropriate model choices.

In Li et al. [2001a], a multiview based face recognition system was proposed to recognize faces from videos with large pose variations. To address the challenging pose issue, the concept of an *identity surface* that captures joint spatial and temporal information was used. An identity surface is a hypersurface formed by projecting all the images of one individual onto the discriminating feature space parameterized on head pose (Figure 16).[14] To characterize the head pose, two angles, yaw and tilt, are used as basis coordinates in the feature space. As plotted in Figure 16, the other basis coordinates represent discriminating feature patterns of faces; this will be discussed later. Based on recovered pose information, a trajectory of the input feature pattern can be constructed. The trajectories of features from known subjects arranged in the same temporal order can be synthesized on their respective identity surfaces. To recognize a face across views over time, the trajectory for the input face is matched to the trajectories synthesized for the known subjects. This approach can be thought of as a generalized version of face recognition based on single images taken at different poses. Experimental results using twelve training sequences, each containing one subject, and new testing sequences of these subjects were reported. Recognition rates were 100% and 93.9%, using 10 and 2 KDA (kernel discriminant analysis) vectors, respectively.

Other techniques have also been used to construct the discriminating basis in the identity surface: kernel discriminant analysis (KDA) [Mika et al. 1999] was used to compute a nonlinear discriminating basis, and a dynamic face model is used to extract a shape-and-pose-free facial texture pattern. The multiview dynamic face model [Li et al. 2001b] consists of a sparse Point Distribution Model (PDM) [Cootes et al. 1995], a shape-and-pose-free texture model, and an affine geometrical model. The 3D shape vector of a face is estimated from a set of 2D face images in different views using landmark points. Then a face image fitted by the shape model is warped to the mean shape in a frontal view, yielding a shape-and-pose-free texture pattern.[15] When part of a face is invisible in an image due to rotation in depth, the facial texture is recovered from the visible side of the face using the bilateral symmetry of faces. To obtain a low-dimensional statistical model, PCA was applied to the 3D shape patterns and shape-and-pose-free texture patterns separately. To further suppress within-class variations, the shape-and-pose-free texture patterns were further projected into a KDA feature space. Finally, the identity surface can be approximated and constructed from discrete samples at fixed poses using a piece-wise planar model.

4.3. Summary

The availability of video/image sequences gives video-based face recognition a distinct advantage over still-image-based face recognition: the abundance of temporal information. However, the typically low-quality images in video present a significant challenge: the loss of spatial

[14] Notice that this view-based idea has already been explored, for example, in Pentland et al. [1994].

[15] Notice that this procedure is very similar to AAM [Cootes et al. 2001].

information. The key to building a successful video-based system is to use temporal information to compensate for the lost spatial information. For example, a high-resolution frame can in principle be reconstructed from a sequence of low-resolution video frames and used for recognition. A further step is to use the image sequence to reconstruct the 3D shape of the tracked face object via SfM and thus enhance face recognition performance. Finally, a comprehensive approach is to use spatial and temporal information simultaneously for face recognition. This is also supported by related psychological studies.

However, many issues remain for existing systems:

—SfM is a common technique used in computer vision for recovering 3D information from video sequences. However, a major obstacle exists to applying this technique in face recognition: the accuracy of 3D shape recovery. Face images contain smooth, textureless regions and are often acquired under varying illumination,[16] resulting in significant difficulties in accurate recovery of 3D information. The accuracy issue may not be very important for face detection, but it is for face recognition, which must differentiate the 3D shapes of similar objects. One possible solution is the complementary use of shape-from-shading, which can utilize the illumination information. A recent paper on using flow-based SfM techniques for face modeling is A. K. R. Chowdhury, and R. Chellappa [2003].

—Up to now, the databases used in many systems have been very small, say 20 subjects. This is partially due to the tremendous amount of storage space needed for video sequences. Fortunately, relatively large video databases exist, for example, the XM2TV database [Messer et al. 1999], the BANCA database [Bailly-Bailliere et al. 2003], and the addition of video into the FERET and FRVT2002 databases. However, large-scale systematic evaluations are still lacking.

—Although we argue that it is best to use both temporal and spatial information for face recognition, existing spatiotemporal methods have not yet shown their full potential. We believe that these types of methods deserve further investigation.

During the past 8 years, recognition of human behavior has been actively studied: facial expression recognition, hand gesture recognition, activity recognition, etc. As pointed out earlier, descriptions of human behavior are useful and are easier to obtain than recognition of faces. Often they provide complementary information for face recognition or additional cues useful for identification. In principle, both gender classification and facial expression recognition can assist in the classification of identity. For recent reviews on facial expression recognition, see Donato et al. [1999] and Pantic and Rothkrantz [2000]. We also believe that analysis of body movements such as gait or hand gestures can help in person recognition.

5. EVALUATION OF FACE RECOGNITION SYSTEMS

Given the numerous theories and techniques that are applicable to face recognition, it is clear that evaluation and benchmarking of these algorithms is crucial. Previous work on the evaluation of OCR and fingerprint classification systems provided insights into how the evaluation of algorithms and systems can be performed efficiently. One of the most important facts learned in these evaluations is that large sets of test images are essential for adequate evaluation. It is also extremely important that the samples be statistically as similar as possible to the images that arise in the application being considered. Scoring should be done in a way that reflects the costs of errors in recognition. Reject-error behavior should be studied, not just forced recognition.

In planning an evaluation, it is important to keep in mind that the operation

[16]Stereo is less sensitive to illumination change but still has difficulty in handling textureless regions.

of a pattern recognition system is statistical, with measurable distributions of success and failure. These distributions are very application-dependent, and no theory seems to exist that can predict them for new applications. This strongly suggests that an evaluation should be based as closely as possible on a specific application.

During the past 5 years, several large, publicly available face databases have been collected and corresponding testing protocols have been designed. The series of FERET evaluations [Phillips et al. 2000b, 1998; Rizvi et al. 1998][17] attracted nine institutions and companies to participate. They were succeeded by the series of FRVT vendor tests. We describe here the most important face databases and their associated evaluation methods, including the XM2VTS and BANCA [Bailly-Bailliere et al. 2003] database.

5.1. The FERET Protocol

Until recently, there did not exist a common FRT evaluation protocol that included large databases and standard evaluation methods. This made it difficult to assess the status of FRT for real applications, even though many existing systems reported almost perfect performance on small databases.

The first FERET evaluation test was administered in August 1994 [Phillips et al. 1998b]. This evaluation established a baseline for face recognition algorithms, and was designed to measure performance of algorithms that could automatically locate, normalize, and identify faces. This evaluation consisted of three tests, each with a different gallery and probe set. (A gallery is a set of known individuals, while a probe is a set of unknown faces presented for recognition.) The first test measured identification performance from a gallery of 316 individuals with one image per person; the second was a false-alarm test; and the third measured the effects of pose changes on performance. The second FERET evaluation was administered in March 1995; it consisted of a single test that measured identification performance from a gallery of 817 individuals, and included 463 duplicates in the probe set [Phillips et al. 1998b]. (A duplicate is a probe for which the corresponding gallery image was taken on a different day; there were only 60 duplicates in the Aug94 evaluation.) The third and last evaluation (Sep96) was administered in September 1996 and March 1997.

5.1.1. Database. Currently, the FERET database is the only large database that is generally available to researchers without charge. The images in the database were initially acquired with a 35-mm camera and then digitized.

The images were collected in 15 sessions between August 1993 and July 1996. Each session lasted 1 or 2 days, and the location and setup did not change during the session. Sets of 5 to 11 images of each individual were acquired under relatively unconstrained conditions; see Figure 17. They included two frontal views; in the first of these (**fa**) a neutral facial expression was requested and in the second (**fb**) a different facial expression was requested (these requests were not always honored). For 200 individuals, a third frontal view was taken using a different camera and different lighting; this is referred to as the ***fc*** *image*. The remaining images were nonfrontal and included right and left profiles, right and left quarter profiles, and right and left half profiles. The FERET database consists of 1564 sets of images (1199 original sets and 365 duplicate sets)—a total of 14,126 images. A development set of 503 sets of images were released to researchers; the remaining images were sequestered for independent evaluation. In late 2000 the entire FERET database was released along with the Sep96 evaluation protocols, evaluation scoring code, and baseline PCA algorithms.

5.1.2. Evaluation. For details of the three FERET evaluations, see Phillips et al. [2000, 1998b] and Rizvi et al. [1998]. The results of the most recent FERET

[17]http://www.itl.nist.gov/iad/humanid/feret/.

Fig. 17. Images from the FERET dataset; these images are of size 384 × 256.

evaluation (Sep96) will be briefly reviewed here. Because the entire FERET data set has been released, the Sep96 protocol provides a good benchmark for performance of new algorithms. For the Sep96 evaluation, there was a primary gallery consisting of one frontal image (**fa**) per person for 1196 individuals. This was the core gallery used to measure performance for the following four different probe sets:

—**fb** probes—gallery and probe images of an individual taken on the same day with the same lighting (1195 probes);

—**fc** probes—gallery and probe images of an individual taken on the same day with different lighting (194 probes);

—Dup I probes—gallery and probe images of an individual taken on different days—duplicate images (722 probes); and

—Dup II probes—gallery and probe images of an individual taken over a year apart (the gallery consisted of 894 images; 234 probes).

Performance was measured using two basic methods. The first measured identification performance, where the primary performance statistic is the percentage of probes that are correctly identified by the algorithm. The second measured verification performance, where the primary performance measure is the equal error rate between the probability of false alarm and the probability of correct verification. (A more complete method of reporting identification performance is a cumulative match characteristic; for verification performance it is a receiver operating characteristic (ROC).)

The Sep96 evaluation tested the following 10 algorithms:

—an algorithm from Excalibur Corporation (Carlsbad, CA)(Sept. 1996);

—two algorithms from MIT Media Laboratory (Sept. 1996) [Moghaddam et al. 1996; Turk and Pentland 1991];

—three linear discriminant analysis-based algorithms from Michigan State University [Swets and Weng 1996b] (Sept. 1996) and the University of Maryland [Etemad and Chellappa 1997; Zhao et al. 1998] (Sept. 1996 and March 1997);

—a gray-scale projection algorithm from Rutgers University [Wilder 1994] (Sept. 1996);

—an Elastic Graph Matching algorithm from the University of Southern California [Okada et al. 1998; Wiskott et al. 1997] (March 1997);

—a baseline PCA algorithm [Moon and Phillips 2001; Turk and Pentland 1991]; and

—a baseline normalized correlation matching algorithm.

Three of the algorithms performed very well: probabilistic eigenface from MIT [Moghaddam et al. 1996], subspace LDA from UMD [Zhao et al. 1998, 1999], and Elastic Graph Matching from USC [Wiskott et al. 1997].

A number of lessons were learned from the FERET evaluations. The first is that performance depends on the probe category and there is a difference between best and average algorithm performance.

Another lesson is that the scenario has an impact on performance. For

identification, on the **fb** and duplicate probes, the USC scores were 94% and 59%, and the UMD scores were 96% and 47%. However, for verification, the equal error rates were 2% and 14% for USC, and 1% and 12% for UMD.

5.1.3. Summary. The availability of the FERET database and evaluation technology has had a significant impact on progress in the development of face recognition algorithms. The series of tests has allowed advances in algorithm development to be quantified—for example, the performance improvements in the MIT algorithms between March 1995 and September 1996, and in the UMD algorithms between September 1996 and March 1997.

Another important contribution of the FERET evaluations is the identification of areas for future research. In general the test results revealed three major problem areas: recognizing duplicates, recognizing people under illumination variations, and recognizing them under pose variations.

5.1.4. FRVT 2000. The Sep96 FERET evaluation measured performance on prototype laboratory systems. After March 1997 there was rapid advancement in the development of commercial face recognition systems. This advancement represented both a maturing of face recognition technology, and the development of the supporting system and infrastructure necessary to create commercial off-the-shelf (COTS) systems. By the beginning of 2000, COTS face recognition systems were readily available.

To assess the state of the art in COTS face recognition systems the Face Recognition Vendor Test (FRVT) 2000[18] was organized [Blackburn et al. 2001]. FRVT 2000 was a technology evaluation that used the Sep96 evaluation protocol, but was significantly more demanding than the Sep96 FERET evaluation.

Participation in FRVT 2000 was restricted to COTS systems, with companies from Australia, Germany, and the United States participating. The five companies evaluated were Banque-Tec International Pty. Ltd., C-VIS Computer Vision und Automation GmbH, Miros, Inc., Lau Technologies, and Visionics Corporation.

A greater variety of imagery was used in FRVT 2000 than in the FERET evaluations. FRVT 2000 reported results in eight general categories: compression, distance, expression, illumination, media, pose, resolution, and temporal. There was no common gallery across all eight categories; the sizes of the galleries and probe sets varied from category to category.

We briefly summarize the results of FRVT 2000. Full details can be found in [Blackburn et al. 2001], and include identification and verification performance statistics. The media experiments showed that changes in media do not adversely affect performance. Images of a person were taken simultaneously on conventional film and on digital media. The compression experiments showed that compression does not adversely affect performance. Probe images compressed up to 40:1 did not reduce recognition rates. The compression algorithm was JPEG.

FRVT 2000 also examined the effect of pose angle on performance. The results show that pose does not significantly affect performance up to $\pm 25^\circ$, but that performance is significantly affected when the pose angle reaches $\pm 40^\circ$.

In the illumination category, two key effects were investigated. The first was lighting change indoors. This was equivalent to the **fc** probes in FERET. For the best system in this category, the indoor change of lighting did not significantly affect performance. A second experiment tested recognition with an indoor gallery and an outdoor probe set. Moving from indoor to outdoor lighting significantly affected performance, with the best system achieving an identification rate of only 0.55.

The temporal category is equivalent to the duplicate probes in FERET. To compare progress since FERET, dup I and dup II scores were reported. For FRVT 2000 the dup I identification rate was 0.63

[18] http://www.frvt.org.

compared with 0.58 for FERET. The corresponding rates for dup II were 0.64 for FRVT 2000 and 0.52 for FERET. These results showed that there was algorithmic progress between the FERET and FRVT 2000 evaluations. FRVT 2000 showed that two common concerns, the effects of compression and recording media, do not affect performance. It also showed that future areas of interest continue to be duplicates, pose variations, and illumination variations generated when comparing indoor images with outdoor images.

5.1.5. FRVT 2002. The Face Recognition Vendor Test (FRVT) 2002 [Phillips et al. 2003][18] was a large-scale evaluation of automatic face recognition technology. The primary objective of FRVT 2002 was to provide performance measures for assessing the ability of automatic face recognition systems to meet real-world requirements. Ten participants were evaluated under the direct supervision of the FRVT 2002 organizers in July and August 2002.

The heart of the FRVT 2002 was the high computational intensity test (HCInt). The HCInt consisted of 121,589 operational images of 37,437 people. The images were provided from the U.S. Department of State's Mexican nonimmigrant Visa archive. From this data, real-world performance figures on a very large data set were computed. Performance statistics were computed for verification, identification, and watch list tasks.

FRVT 2002 results showed that normal changes in indoor lighting do not significantly affect performance of the top systems. Approximately the same performance results were obtained using two indoor data sets, with different lighting, in FRVT 2002. In both experiments, the best performer had a 90% verification rate at a false accept rate of 1%. On comparable experiments conducted 2 years earlier in FRVT 2000, the results of FRVT 2002 indicated that there has been a 50% reduction in error rates. For the best face recognition systems, the recognition rate for faces captured outdoors, at a false accept rate of 1%, was only 50%. Thus, face recognition from outdoor imagery remains a research challenge area.

A very important question for real-world applications is the rate of decrease in performance as time increases between the acquisition of the database of images and new images presented to a system. FRVT 2002 found that for the top systems, performance degraded at approximately 5% per year.

One open question in face recognition is: how does database and watch list size effect performance? Because of the large number of people and images in the FRVT 2002 data set, FRVT 2002 reported the first large-scale results on this question. For the best system, the top-rank identification rate was 85% on a database of 800 people, 83% on a database of 1,600, and 73% on a database of 37,437. For every doubling of database size, performance decreases by two to three overall percentage points. More generally, identification performance decreases linearly in the logarithm of the database size.

Previous evaluations have reported face recognition performance as a function of imaging properties. For example, previous reports compared the differences in performance when using indoor versus outdoor images, or frontal versus nonfrontal images. FRVT 2002, for the first time, examined the effects of demographics on performance. Two major effects were found. First, recognition rates for males were higher than females. For the top systems, identification rates for males were 6% to 9% points higher than that of females. For the best system, identification performance on males was 78% and for females it was 79%. Second, recognition rates for older people were higher than for younger people. For 18- to 22-year-olds the average identification rate for the top systems was 62%, and for 38- to 42-year-olds it was 74%. For every 10-year increase in age, performance increased on the average by approximately 5% through age 63.

FRVT 2002 looked at two of these new techniques. The first was the three-dimensional morphable models technique

of Blanz and Vetter [1999]. Morphable models are a technique for improving recognition of nonfrontal images. FRVT 2002 found that Blanz and Vetter's technique significantly increased recognition performance. The second technique is recognition from video sequences. Using FRVT 2002 data, recognition performance using video sequences was the same as the performance using still images.

In summary, the key lessons learned in FRVT 2002 were: (1) given reasonable controlled indoor lighting, the current state of the art in face recognition is 90% verification at a 1% false accept rate. (2) Face Recognition in outdoor images is a research problem. (3) The use of morphable models can significantly improve nonfrontal face recognition. (3) Identification performance decreases linearly in the logarithm of the size of the gallery. (4) In face recognition applications, accommodations should be made for demographic information since characteristics such as age and sex can significantly affect performance.

5.2. The XM2VTS Protocol

Multimodal methods[19] are a very promising approach to user-friendly (hence acceptable), highly secure personal verification. Recognition and verification systems need training; the larger the training set, the better the performance achieved. The volume of data required for training a multimodal system based on analysis of video and audio signals is on the order of TBytes; technology that allows manipulation and effective use of such volumes of data has only recently become available in the form of digital video. The XM2VTS multimodal database [Messer et al. 1999] contains four recordings of 295 subjects taken over a period of 4 months. Each recording contains a speaking head shot and a rotating head shot. Available data from this database include high-quality color images, 32-kHz 16-bit sound files, video sequences, and a 3D model.

[19]http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/.

The XM2VTS database is an expansion of the earlier M2VTS database [Pigeon and Vandendorpe 1999]. The M2VTS project (Multimodal Verification for Teleservices and Security Applications), a European ACTS (Advanced Communications Technologies and Services) project, deals with access control by multimodal identification of human faces. The goal of the project was to improve recognition performance by combining the modalities of face and voice. The M2VTS database contained five shots of each of 37 subjects. During each shot, the subjects were asked to count from "0" to "9" in their native language (most of the subjects were French-speaking) and rotate their heads from 0° to -90°, back to 0°, and then to $+90^\circ$. They were then asked to rotate their heads again with their glasses off, if they wore any. Three subsequences were extracted from these video sequences: voice sequences, motion sequences, and glasses-off motion sequences. The voice sequences can be used for speech verification, frontal view face recognition, and speech/lips correlation analysis. The other two sequences are intended for face recognition only.

It was found that the subjects were relatively difficult to recognize in the fifth shot because it varied significantly in face/voice/camera setup from the other shots. Several experiments have been conducted using the first four shots with the goals of investigating

—text-dependent speaker verification from speech,
—text-independent speaker verification from speech,
—facial feature extraction and tracking from moving images,
—verification from an overall frontal view,
—verification from lip shape,
—verification from depth information (obtained using structured light),
—verification from a profile, and
—synchronization of speech and lip movement.

5.2.1. Database. The XM2VTS database differs from the M2VTS database primarily in the number of subjects (295 rather than 37). The M2VTS database contains five shots of each subject taken at sessions over a period of 3 months; the XM2VTS database contains eight shots of each subject taken at four sessions over a period of 4 months (so that each session contains two repetitions of the sequence). The XM2VTS database was acquired using a Sony VX1000E digital camcorder and a DHR1000UX digital VCR.

In the XM2VTS database, the first shot is a speaking head shot. Each subject, who wore a clip-on microphone, was asked to read three sentences that were written on a board positioned just below the camera. The subjects were asked to read the three simple sentences twice at their normal pace and to pause briefly at the end of each sentence.

The second shot is a rotating head sequence. Each subject was asked to rotate his/her head to the left, to the right, up, and down, and finally to return to the center. The subjects were told that a full profile was required and were asked to repeat the entire sequence twice. The same sequence was used in all four sessions.

An additional dataset containing a 3D model of each subject's head was acquired during each session using a high-precision stereo-based 3D camera developed by the Turing Institute.[20]

5.2.2. Evaluation. The M2VTS Lausanne protocol was designed to evaluate the performance of vision- and speech-based person authentication systems on the XM2VTS database. This protocol was defined for the task of verification. The features of the observed person are compared with stored features corresponding to the claimed identity, and the system decides whether the identity claim is true or false on the basis of a similarity score. The subjects whose features are stored in the system's database are called *clients*, whereas persons claiming a false identity are called *imposters*.

The database is divided into three parts: a training set, an evaluation set, and a test set. The training set is used to build client models. The evaluation set is used to compute client and imposter scores. On the basis of these scores, a threshold is chosen that determines whether a person is accepted or rejected. In multimodal classification, the evaluation set can also be used to optimally combine the outputs of several classifiers. The test set is selected to simulate a real authentication scenario. 295 subjects were randomly divided into 200 clients, 25 evaluation imposters, and 70 test imposters. Two different evaluation configurations were used with different distributions of client training and client evaluation data. For more details, see Messer et al. [1999].

In order to collect face verification results on this database using the Lausanne protocol, a contest was organized in conjuction with ICPR 2000 (the International Conference on Pattern Recognition). There were twelve algorithms from four partipicants in this contest [Matas et al. 2000]: an EBGM algorithm from IDIAP (Daller Molle Institute for Perceptual Artificial Intelligence), a slightly modified EBGM algorithm from Aristotle University of Thessaloniki, a FND-based (Fractal Neighbor Distance) algorithm from the University of Sydney, and eight variants of LDA algorithms and one SVM algorithm from the University of Surrey. The performance measures of a verification system are the false acceptance rate (FA) and the false rejection rate (FR). Both FA and FR are influenced by an acceptance threshold. According to the Lausanne protocol, the threshold is set to satisfy certain performance levels on the evaluation set. The same threshold is applied to the test data and FA and FR on the test data are computed. The best results of FA and FR on the test data (FA/FR: 2.3%/2.5% and 1.2%/1.0% for evaulation configurations I and II, respectively) were obtained using an LDA algorithm with a non-Euclidean metric (University of Surrey) when the threshold was set so that FA was equal to FB on the evaulation result. This result seems to concur with the

[20]Turing Institute Web address: http://www.turing.gla.ac.uk/.

equal error rates reported in the FERET protocol. In addition, FA and FR on the test data were reported when the threshold was set set so that FA or FB was zero on the evaulation result. For more details on the results, see Matas et al. [2000].

5.2.3. Summary. The results of the M2VTS/XM2VTS projects can be used for a broad range of applications. In the telecommunication field, the results should have a direct impact on network services where security of information and access will become increasingly important. (Telephone fraud in the U.S. has been estimated to cost several billion dollars a year.)

6. TWO ISSUES IN FACE RECOGNITION: ILLUMINATION AND POSE VARIATION

In this section, we discuss two important issues that are related to face recognition. The best face recognition techniques reviewed in Section 3 were successful in terms of their recognition performance on large databases in well-controlled environments. However, face recognition in an uncontrolled environment is still very challenging. For example, the FERET evaluations and FRVTs revealed that there are at least two major challenges: the illumination variation problem and the pose variation problem. Though many existing systems build in some sort of performance invariance by applying preprocessing methods such as histogram equalization or pose learning, significant illumination or pose change can cause serious performance degradation. In addition, face images can be partially occluded, or the system may need to recognize a person from an image in the database that was acquired some time ago (referred to as the *duplicate* problem in the FERET tests).

These problems are unavoidable when face images are acquired in an uncontrolled, uncooperative environment, as in surveillance video clips. It is beyond the scope of this paper to discuss all these issues and possible solutions. In this section we discuss only two well-defined problems and review approaches to solving them. Pros and cons of these approaches are pointed out so an appropriate approach can be applied to a specific task. The majority of the methods reviewed here are *generative* approaches that can synthesize virtual views under desired illumination and viewing conditions. Many of the reviewed methods have not yet been applied to the task of face recognition, at least not on large databases.[21] This may be for several reasons; some methods may need many sample images per person, pixel-wise accurate alignment of images, or high-quality images for reconstruction; or they may be computationally too expensive to apply to recognition tasks that process thousands of images in near-real-time.

To facilitate discussion and analysis, we adopt a varying-albedo Lambertian reflectance model that relates the image I of an object to the object (p, q) [Horn and Brooks 1989]:

$$I = \rho \frac{1 + pP_s + qQ_s}{\sqrt{1+p^2+q^2}\sqrt{1+P_s^2+Q_s^2}}, \quad (6)$$

where (p, q), ρ are the partial derivatives and varying albedo of the object, respectively. $(P_s, Q_s, -1)$ represents a single distant light source. The light source can also be represented by the illuminant slant and tilt angles; *slant* α is the angle between the opposite lighting direction and the positive z-axis, and *tilt* τ is the angle between the opposite lighting direction and the x-z plane. These angles are related to P_s and Q_s by $P_s = \tan\alpha\cos\tau$, $Q_s = \tan\alpha\sin\tau$. To simplify the notation, we replace the constant $\sqrt{1+P_s^2+Q_s^2}$ by K. For easier analysis, we assume that frontal face objects are bilaterally symmetric about the vertical midlines of the faces.

[21]One exception is a recent report [Blanz and Vetter 2003] where faces were represented using 4448 images from the CMU-PIE databases and 1940 images from the FERET database.

Fig. 18. In each row, the same face appears differently under different illuminations (from the Yale face database).

6.1. The Illumination Problem in Face Recognition

The illumination problem is illustrated in Figure 18, where the same face appears different due to a change in lighting. The changes induced by illumination are often larger than the differences between individuals, causing systems based on comparing images to misclassify input images. This was experimentally observed in Adini et al. [1997] using a dataset of 25 individuals.

In Zhao [1999], an analysis was carried out of how illumination variation changes the eigen-subspace projection coefficients of images under the assumption of a Lambertian surface. Consider the basic expression for the subspace decomposition of a face image I: $I \simeq I_A + \sum_{i=1}^{m} a_i \Phi_i$, where I_A is the average image, Φ_i are the eigenimages, and a_i are the projection coefficients. Assume that for a particular individual we have a prototype image I_p that is a normally lighted frontal view ($P_s = 0, Q_s = 0$ in Equation (6)) in the database, and we want to match it against a new image $\tilde{I}$ of the same class under lighting $(P_s, Q_s, -1)$. The corresponding subspace projection coefficient vectors $\mathbf{a} = [a_1, a_2, \ldots, a_m]^T$ (for I_p) and $\tilde{\mathbf{a}} = [\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_m]^T$ (for $\tilde{I}$) are computed as follows:

$$\begin{aligned} a_i &= I_p \odot \Phi_i - I_A \odot \Phi_i, \\ \tilde{a}_i &= \tilde{I} \odot \Phi_i - I_A \odot \Phi_i, \end{aligned} \tag{7}$$

where $\odot$ denotes the sum of all element-wise products of two matrices (vectors). If we divide the images and the eigenimages into two halves, for example, left and right, we have

$$\begin{aligned} a_i &= I_p^L \odot \Phi_i^L + I_p^R \odot \Phi_i^R - I_A \odot \Phi_i, \\ \tilde{a}_i &= \tilde{I}^L \odot \Phi_i^L + \tilde{I}^R \odot \Phi_i^R - I_A \odot \Phi_i. \end{aligned} \tag{8}$$

Based on Equation (6), the symmetric property of eigenimages and face objects, we have

$$\begin{aligned} a_i &= 2I_p^L[x,y] \odot \Phi_i^L[x,y] - I_A \odot \Phi_i, \\ \tilde{a}_i &= \left(\frac{2}{K}\right)(I_p^L[x,y] + I_p^L[x,y]q^L[x,y]Q_s) \\ &\quad \odot \Phi_i^L[x,y] - I_A \odot \Phi_i, \end{aligned} \tag{9}$$

leading to the following relation:

$$\begin{aligned} \tilde{\mathbf{a}} &= \left(\frac{1}{K}\mathbf{a}\right) + \frac{Q_s}{K}\left[f_1^a, f_2^a, \ldots, f_m^a\right]^T \\ &\quad - \frac{K-1}{K}\mathbf{a}_A., \end{aligned} \tag{10}$$

where $f_i^a = 2(I_p^L[x,y]q^L[x,y]) \odot \Phi_i^L[x,y]$ and $\mathbf{a}_A$ is the projection coefficient vector of the average image I_A: $[I_A \odot \Phi_1, \ldots, I_A \odot \Phi_m]$. Now let us assume that the training set is extended to include mirror images as in Kirby and Sirovich [1990]. A similar analysis can be carried out, since in such a case the eigenimages are either symmetric (for most leading eigenimages) or antisymmetric.

In general, Equation (11) suggests that a significant illumination change can seriously degrade the performance of

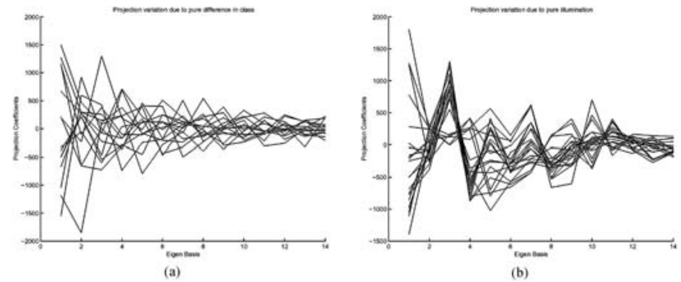

Fig. 19. Changes of projection vectors due to class variation (a) and illumination change (b) is of the same order [Zhao 1999].

subspace-based methods. Figure 19 plots the projection coefficients for the same face under different illuminations ($\alpha \in [0^0, 40^0]$, $\tau \in [0^0, 180^0]$) and compares them against the variations in the projection coefficient vectors due to pure differences in class.

In general, the illumination problem is quite difficult and has received considerable attention in the image understanding literature. In the case of face recognition, many approaches to this problem have been proposed that make use of the domain knowledge that all faces belong to one face class. These approaches can be divided into four types [Zhao 1999]: (1) heuristic methods, for example, discarding the leading principal components; (2) image comparison methods in which appropriate image representations and distance measures are used; (3) class-based methods using multiple images of the same face in a fixed pose but under different lighting conditions; and (4) model-based approaches in which 3D models are employed.

6.1.1. Heuristic Approaches. Many existing systems use heuristic methods to compensate for lighting changes. For example, in Moghaddam and Pentland [1997] simple contrast normalization was used to preprocess the detected faces, while in Sung and Poggio [1997] normalization in intensity was done by first subtracting a best-fit brightness plane and then applying histogram equalization. In the face eigen-subspace domain, it was suggested and later experimentally verified in Belhumeur et al. [1997] that by discarding a few most significant principal components, variations due to lighting can be reduced. The plot in Figure 19(b) also supports this observation. However, in order to maintain system performance for normally illuminated images, while improving performance for images acquired under changes in illumination, it must be assumed that the first three principal components capture only variations due to lighting. Other heuristic methods based on frontal-face symmetry have also been proposed [Zhao 1999].

6.1.2. Image Comparison Approaches. In Adini et al. [1997], approaches based on image comparison using different image representations and distance measures were evaluated. The image representations used were edge maps, derivatives of the gray level, images filtered with 2D Gabor-like functions, and a representation that combines a log function of the intensity with these representations. The distance measures used were point-wise distance, regional distance, affine-GL (gray level) distance, local affine-GL distance, and log point-wise distance. For more details about these methods and about the evaluation database, see Adini et al. [1997]. It was concluded that none of these representations alone can overcome the image variations due to illumination.

A recently proposed image comparison method Jacobs et al. [1998] used a new measure robust to illumination change. The rationale for develop such a method of directly comparing images is the potential difficulty of building a complete representation of an object's possible images as suggested in [Belhumeur and Kriegman 1997]. The authors argued that it is not clear whether it is possible to construct the complete representation using a small number of training images taken under uncontrolled viewing conditions and containing multiple light sources. It was shown that given two images of an object with unknown structure and albedo, there is always a large family of solutions. Even in the case of given light sources, only two out of three independent components of the Hessian of the surface can be determined. Instead, the authors argued that the ratio of two images of the same object is simpler than if the images are from different objects. Based on this observation, the complexity of the ratio of two aligned images was proposed as the similarity measure. More specifically, we have

$$\frac{I_1}{I_2} = \left(\frac{K_2}{K_1}\right)\left(\frac{1+p_I P_{s,1}+q_I Q_{s,1}}{1+p_I P_{s,2}+q_I Q_{s,2}}\right) \quad (11)$$

for images of the same object, and

$$\frac{I_1}{J_2} = \left(\frac{K_2}{K_1}\right)\left(\frac{\rho_I}{\rho_J}\right)\left(\frac{1+p_I P_{s,1}+q_I Q_{s,1}}{1+p_J P_{s,2}+q_J Q_{s,2}}\right) \times\sqrt{\frac{1+p_J^2+q_J^2}{1+p_I^2+q_I^2}} \quad (12)$$

for images of different objects. They chose the integral of the magnitude of the gradient of the function (ratio image) as the measure of complexity and proposed the following symmetric similarity measure:

$$d_G(I,J) = \int\int \min(I,J)\left\|\Delta\left(\frac{I}{J}\right)\right\| \left\|\Delta\frac{J}{I}\right\| dx\,dy. \quad (13)$$

They noticed the similarity between this measure and the measure that simply compares the edges. It is also clear that the measure is not strictly illumination-invariant because it changes for a pair of images of the same object when the illumination changes. Experiments on face recognition showed improved performance over eigenfaces, which were somewhat worse than the illumination cone-based method [Georghiades et al. 1998] on the same set of data.

6.1.3. Class-Based Approaches. Under the assumptions of Lambertian surfaces and no shadowing, a 3D linear illumination subspace for a person was constructed in Belhumeur and Kriegman [1997], Hallinan [1994], Murase and Nayar [1995], Ricklin-Raviv and Shashua [1999], and Shashua [1994] for a fixed viewpoint, using three aligned faces/images acquired under different lighting conditions. Under ideal assumptions, recognition based on this subspace is illumination-invariant. More recently, an illumination cone has been proposed as an effective method of handling illumination variations, including shadowing and multiple light sources [Belhumeur and Kriegman 1997; Georghiades et al. 1998]. This method is an extension of the 3D linear subspace method [Hallinan 1994; Shashua 1994] and has the same drawback, requiring at least three aligned training images acquired under different lighting conditions per person. A more detailed review of this approach and its extension to handle the combined illumination and pose problem will be presented in Section 6.2.

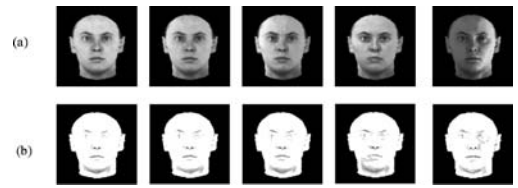

Fig. 20. Testing the invariance of the quotient image (Q-image) to varying illumination. (a) Original images of a novel face taken under five different illuminations. (b) The Q-images corresponding to the novel images, computed with respect to the bootstrap set of ten objects [Riklin-Raviv and Shashua 1999]. (Courtesy of T. Riklin-Raviv and A. Shashua.)

More recently, a method based on *quotient images* was introduced [Riklin-Raviv and Shashua 1999]. Like other class-based methods, this method assumes that the faces of different individuals have the same shape and different textures. Given two objects **a**, **b**, the quotient image Q is defined to be the ratio of their albedo functions ρ_a/ρ_b, and hence is illumination-invariant. Once Q is computed, the entire illumination space of object **a** can be generated by Q and a linear illumination subspace constructed from three images of object **b**. To make this basic idea work in practice, a training set (called the *bootstrap set* in the paper) is needed that consists of images of N objects under various lighting conditions, and the quotient image of a novel object **y** is defined relative to the average object of the bootstrap set. More specifically, the bootstrap set consists of $3N$ images taken from three fixed, linearly independent light sources s_1, s_2, and s_3 that are not known. Under this assumption, any light source s can be expressed as a linear combination of the s_i: $s = x_1 s_1 + x_2 s_2 + x_3 s_3$. The authors further defined the normalized albedo function ρ of the bootstrap set as the squared sum of the ρ_i, where ρ_i is the albedo function of object i. An interesting energy/cost function is defined that is quite different from the traditional bilinear form. Let $A_1, A_2, \ldots, A_N$ be $m \times 3$ matrices whose columns are images of object i (from the bootstrap set) that contain the same m pixels; then the bilinear energy/cost function [Freeman and Tenenbaum 2000] for an image y_s of object **y** under illumination s is

$$\left(y_s - \sum_{i=1}^{N} \alpha_i A_i x\right)^2, \quad (14)$$

which is a bilinear problem in the N unknowns α_i and 3 unknowns x. For comparison, the proposed energy function is

$$\sum_{i=1}^{N} (\alpha_i y_s - A_i x)^2. \quad (15)$$

This formation of the energy function is a major reason why the quotient image method works better than "reconstruction" methods based on Equation (14) in terms of smaller size of the bootstrap set and less requirement for pixel-wise image alignment. As pointed out by the authors, another factor contributing to the success of using only a small bootstrap set is that the albedo functions occupy only a small subspace. Figure 20 demonstrates the invariance of the quotient image against change in illumination conditions; the image synthesis results are shown in Figure 21.

6.1.4. Model-Based Approaches. In model-based approaches, a 3D face model is used to synthesize the virtual image from a given image under desired illumination conditions. When the 3D model is unknown, recovering the shape from

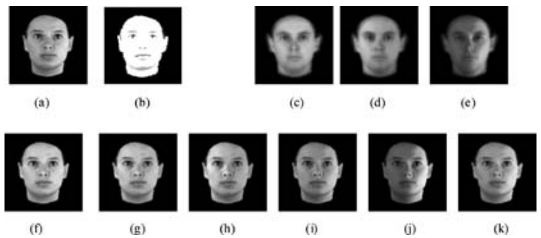

Fig. 21. Image synthesis example. Original image (a) and its quotient image (b) from the $N = 10$ bootstrap set. The quotient image is generated relative to the average object of the bootstrap set, shown in (c), (d), and (e). Images (f) through (k) are synthetic images created from (b) and (c), (d), (e) [Riklin-Raviv and Shashua 1999]. (Courtesy of T. Riklin-Raviv and A. Shashua.)

the images *accurately* is difficult without using any priors. Shape-from-shading (SFS) can be used if only one image is available; stereo or structure from motion can be used when multiple images of the same object are available.

Fortunately, for face recognition the differences in the 3D shapes of different face objects are not dramatic. This is especially true after the images are aligned and normalized. Recall that this assumption was used in the class-based methods reviewed above. Using a statistical representation of the 3D heads, PCA was suggested as a tool for solving the parametric SFS problem [Atick et al. 1996]. An eigenhead approximation of a 3D head was obtained after training on about 300 laser-scanned range images of real human heads. The ill-posed SFS problem is thereby transformed into a parametric problem. The authors also demonstrated that such a representation helps to determine the light source. For a new face image, its 3D head can be approximated as a linear combination of eigenheads and then used to determine the light source. Using this complete 3D model, any virtual view of the face image can be generated. A major drawback of this approach is the assumption of *constant* albedo. This assumption does not hold for most real face images, even though it is the most common assumption used in SFS algorithms.

To address the issue of varying albedo, a *direct* 2D-to-2D approach was proposed based on the assumption that front-view faces are symmetric and making use of a generic 3D model [Zhao et al. 1999]. Recall that a prototype image I_p is a frontal view with $P_s = 0$, $Q_s = 0$. Substituting this into Equation (6), we have

$$I_p[x, y] = \rho \frac{1}{\sqrt{1 + p^2 + q^2}}. \tag{16}$$

Comparing Equations (6) and (16), we obtain

$$I_p[x, y] = \frac{K}{2(1 + q Q_s)}(I[x, y] + I[-x, y]). \tag{17}$$

This simple equation relates the prototype image I_p to $I[x, y] + I[x, -y]$, which is already available. The two advantages of this approach are: (1) there is no need to recover the varying albedo $\rho[x, y]$; (2) there is no need to recover the full shape gradients (p, q); q can be approximated by a value derived from a generic 3D shape. As part of the proposed automatic method, a model-based light source identification method was also proposed to improve existing source-from-shading algorithms. Figure 22 shows some comparisons between rendered images obtained using this method and using a

Fig. 22. Image rendering comparison. The original images are shown in the first column. The second column shows prototype images rendered using the local SFS algorithm [Tsai and Shah 1994]. Prototype images rendered using symmetric SFS are shown in the third column. Finally, the fourth column shows real images that are close to the prototype images [Zhao and Chellappa 2000].

local SFS algorithm [Tsai and Shah 1994]. Using the Yale and Weizmann databases (Table V), significant performance improvements were reported when the prototype images were used in a subspace LDA system in place of the original input images [Zhao et al. 1999]. In these experiments, the gallery set contained about 500 images from various databases and the probe set contained 60 images from the Yale database and 96 images from the Weizmann database.

Recently, a general method of approximating Lambertian reflectance using second-order spherical harmonics has been reported [Basri and Jacobs 2001]. Assuming Lambertian objects under distant, isotropic lightng, the authors were able to show that the set of all reflectance functions can be approximated using the surface spherical harmonic expansion. Specifically, they have proved that using a second-order (nine harmonics, i.e., nine-dimensional 9D-space) approximation, the accuracy for any light function exceeds 97.97%. They then extended this analysis to image formation, which is a much more difficult problem due to possible occlusion, shape, and albedo variations. As indicated by the authors, worst-case image approximation can be arbitrarily bad, but most cases are good. Using their method, an image can be decomposed into so-called *harmonic images*, which are produced when the object is illuminated by harmonic functions. The nine harmonic images of a face are plotted in Figure 23. An interesting comparison was made between the proposed method and the 3D linear illumination subspace methods [Hallinan 1994; Shashua 1994]; the 3D linear methods are just first-order harmonic approximations without the DC components.

Assuming precomputed object pose and known color albedo/texture, the authors reported an 86% correct recognition rate when applying this technique to the task of face recognition using a probe set of 10 people and a gallery set of 42 people.

6.2. The Pose Problem in Face Recognition

It is not surprising that the performance of face recognition systems drops significantly when large pose variations are present, in the input images. This difficulty was documented in the FERET and FRVT test reports [Blackburn et al. 2001; Phillips et al. 2002b, 2003], and was suggested as a major research issue. When illumination variation is also present, the task of face recognition becomes even more difficult. Here we focus on the out-of-plane rotation problem, since in-plane rotation is a pure 2D problem and can be solved much more easily.

Table V. Internet Resources for Research and Databases

Research pointers	
Face recognition homepage	www.cs.rug.nl/∼peterkr/FACE/frhp.html
Face detection homepage	home.t-online.de/home/Robert.Frischholz/face.htm
Facial analysis homepage	mambo.ucsc.edu/psl/fanl.html
Facial animiation homepage	mambo.ucsc.edu/psl/fan.html
Face databases	
FERET database	http://www.itl.nist.gov/iad/humanid/feret/
XM2TVS database	http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/
UT Dallas database	http://www.utdallas.edu/dept/bbs/FACULTY_PAGES/otoole/database.htm
Notre Dame database	http://www.nd.edu/∼cvrl/HID-data.html
MIT face databases	ftp://whitechapel.media.mit.edu/pub/images/
Shimon Edelman's face database	ftp://ftp.wisdom.weizmann.ac.il/pub/FaceBase/
CMU face detection database	www.ius.cs.cmu.edu/IUS/dylan_usr0/har/faces/test/
CMU PIE database	www.ri.cmu.edu/projects/project_418.html
Stirling face database	pics.psych.stir.ac.uk
M2VTS multimodal database	www.tele.ucl.ac.be/M2VTS/
Yale face database	cvc.yale.edu/projects/yalefaces/yalefaces.html
Yale face database B	cvc.yale.edu/projects/yalefacesB/yalefacesB.html
Harvard face database	hrl.harvard.edu/pub/faces
Weizmann face database	www.wisdom.weizmann.ac.il/∼yael/
UMIST face database	images.ee.umist.ac.uk/danny/database.html
Purdue University face database	rvl1.ecn.purdue.edu/∼aleix/aleix_face_DB.html
Olivetti face database	www.cam-orl.co.uk/facedatabase.html
Oulu physics-based face database	www.ee.oulu.fi/research/imag/color/pbfd.html

Fig. 23. The first nine harmonic images of a face object (from left to right, top to bottom) [Basri and Jacobs 2001]. (Courtesy of R. Basri and D. Jacobs.)

Earlier methods focused on constructing invariant features [Wiskott et al. 1997] or synthesizing a prototypical view (frontal view) after a full model is extracted from the input image [Lanitis et al. 1995].[22] Such methods work well for small rotation angles, but they fail when the angle is large, say 60°, causing some important features to be invisible. Most proposed methods are based on using large numbers of multiview samples. This seems to concur with the findings of the psychology community; face perception is believed to be view-independent for small angles, but view-dependent for large angles.

To assess the pose problem more systematically, an attempt has been made to classify pose problems [Zhao 1999; Zhao and Chellappa 2000b]. The basic idea of this analysis is to use a varying-albedo reflectance model (Equation (6)) to synthesize new images in different poses from a real image, thus providing a tool for simulating the pose problem. More specifically, the 2D-to-2D image transformation under 3D pose change has been studied. The drawback of this analysis is the restriction of using a generic 3D model; no deformation of this 3D shape was carried out, though the authors suggested doing so.

Researchers have proposed various methods of handling the rotation problem. They can be divided into three classes [Zhao 1999]: (1) multiview image methods, when multiview database images of each person are available; (2) hybrid methods, when multiview training images are available during training but only one database image per person

[22]One exception is the multiview eigenfaces of Pentland et al. [1994].

is available during recognition; and (3) single-image/shape-based methods where no training is carried out. Akamatsu et al. [1992], Beymer [1993], Georghiades et al. [1999, 2001], and Ullman and Basri [1991] are examples of the first class and Beymer [1995], Beymer and Poggio [1995], Cootes et al. [2000], Maurer and Malsburg [1996a], Sali and Ullman [1998], and Vetter and Poggio [1997] of the second class. Up to now, the second type of approach has been the most popular. The third approach does not seem to have received much attention.

6.2.1. Multiview-Based Approaches. One of the earliest examples of the first class of approaches is the work of Beymer [1993], which used a template-based correlation matching scheme. In this work, pose estimation and face recognition were coupled in an iterative loop. For each hypothesized pose, the input image was aligned to database images corresponding to that pose. The alignment was first carried out via a 2D affine transformation based on three key feature points (eyes and nose), and optical flow was then used to refine the alignment of each template. After this step, the correlation scores of all pairs of matching templates were used for recognition. The main limitations of this method, and other methods belonging to this type of approach, are (1) many different views per person are needed in the database; (2) no lighting variations or facial expressions are allowed; and (3) the computational cost is high, since iterative searching is involved.

More recently, an illumination-cone-based [Belhumeur and Kriegman 1997] image synthesis method [Georghiades et al. 1999] has been proposed to handle both pose and illumination problems in face recognition. It handles illumination variation quite well, but not pose variation. To handle variations due to rotation, it needs to completely resolve the GBR (generalized-bas-relief) ambiguity and then reconstruct the Euclidean 3D shape. Without resolving this ambiguity, images from nonfrontal viewpoints synthesized from a GBR reconstruction will differ from a valid image by an affine warp of the image coordinates.[23] To address GBR ambiguity, the authors proposed exploiting face symmetry (to correct tilt) and the fact that the chin and the forehead are at about the same height (to correct slant), and requiring that the range of heights of the surface be about twice the distance between the eyes (to correct scale) [Georghiades et al. 2001]. They propose a pose- and illumination-invariant face recognition method based on building illumination cones at each pose for each person. Though conceptually this is a good idea, in practice it is too expensive to implement. The authors suggested many ways of speeding up the process, including first subsampling the illumination cone and then approximating the subsampled cone with a 11D linear subspace. Experiments on building illumination cones and on 3D shape reconstruction based on seven training images per class were reported. To visualize illumination-cone based image synthesis, see Figure 24. Figure 25 demonstrates the effectiveness of image synthesis under variable pose and lighting after the GBR ambiguity is resolved. Almost perfect recognition results on ten individuals were reported using nine poses and 45 viewing conditions.

6.2.2. Hybrid Approaches. Numerous algorithms of the second type have been proposed. These methods, which make use of prior class information, are the most successful and practical methods up to now. We review several representative methods here: (1) a view-based eigenface method [Pentland et al. 1994], (2) a graph matching-based method [Wiskott et al. 1997], (3) a linear class-based method [Blanz and Vetter 1999; Vetter and Poggio 1997], (4) a vectorized image representation based method [Beymer 1995; Beymer and Poggio 1995], and (5) a view-based appearance model [Cootes

[23]GBR is a 3D affine transformation with three parameters: scale, slant, and tilt. A weak-perspective imaging model is assumed.

Fig. 24. The process of constructing the illumination cone. (a) The seven training images from Subset 1 (near frontal illumination) in frontal pose. (b) Images corresponding to the columns of $\bar{B}$. (c) Reconstruction up to a GBR transformation. On the left, the surface was rendered with flat shading, that is, the albedo was assumed to be constant across the surface, while on the right the surface was texture-mapped using the first basis image of $\bar{B}$ shown in Figure 24(b). (d) Synthesized images from the illumination cone of the face under novel lighting conditions but fixed pose. Note the large variations in shading and shadowing as compared to the seven training images. (Courtesy of A. Georghiades, P. Belhumeur, and D. Kriegman.)

et al. 2000]. Some of the reviewed methods are very closely related—for example, methods 3, 4, and 5. Despite their popularity, these methods have two common drawbacks: (1) they need many example images to cover the range of possible views; (2) the illumination problem is not explicitly addressed, though in principle it can be handled if images captured under the same pose but different illumination conditions are available.

The popular eigenface approach [Turk and Pentland 1991] to face recognition has been extended to a view-based eigenface method in order to achieve pose-invariant recognition [Pentland et al. 1994]. This method explicitly codes the pose information by constructing an individual eigenface for each pose. More recently, a unified framework called the *bilinear model* was proposed in Freeman and Tenenbaum [2000] that can handle either pure pose variation or pure class variation. (A bilinear example is given in Equation (14) for the illumination problem.)

In Wiskott et al. [1997], a robust face recognition scheme based on EBGM was proposed. The authors assumed a planar surface patch at each feature point (landmark), and learned the transformations

Fig. 25. Synthesized images under variable pose and lighting generated from the training images shown in Figure 24 and 25. (Courtesy of A. Georghiades, P. Belhumeur, and D. Kriegman.)

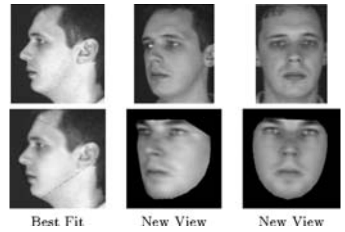

Fig. 26. The best fit to a profile model is projected to the frontal model to predict new views [Cootes et al. 2000]. (Courtesy of T. Cootes, K. Walker, and C. Taylor.)

of "jets" under face rotation. Their results demonstrated substantial improvement in face recognition under rotation. Their method is also fully automatic, including face localization, landmark detection, and flexible graph matching. The drawback of this method is its requirement for accurate landmark localization, which is not an easy task, especially when illumination variations are present.

The image synthesis method in Vetter and Poggio [1997] is based on the assumption of linear 3D object classes and the extension of linearity to images (both shape and texture) that are 2D projections of the 3D objects. It extends the linear shape model (which is very similar to the active shape model of Cootes et al. [1995]) from a representation based on feature points to full images of objects. To implement this method, a correspondence between images of the input object and a reference object is established using optical flow. Correspondences between the reference image and other example images having the same pose are also computed. Finally, the correspondence field for the input image is linearly decomposed into the correspondence fields for the examples. Compared to the parallel deformation scheme in Beymer and Poggio [1995], this method reduces the need to compute correspondences between images of different poses. On the other hand, parallel deformation was able to preserve some peculiarities of texture that are nonlinear and that could be "erased" by linear methods. This method was extended in Sali and Ullman [1998] to include an additive error term for better synthesis. In Blanz and Vetter [1999], a morphable 3D face model consisting of shape and texture was directly matched to single/multiple input images. As a consequence, head orientation, illumination conditions, and other parameters could be free variables subject to optimization.

In Cootes et al. [2000], a view-based statistical method was proposed based on a small number of 2D statistical models (AAM). Unlike most existing methods that can handle only images with rotation angles up to, say 45°, the authors argued that their method can handle even profile views in which many features are invisible. To deal with such large pose variations, they needed sample views at 90° (full profile), 45° (quasiprofile), and 0° (frontal view). A key element that is unique to this method is that for each pose, a different set of features is used. Given a single image of a new person, all the models are used to match the image, and estimation of the pose is achieved by choosing the best fit. To synthesize a new view from the input image, the relationship between models at different views are learned. More specifically, the following steps are needed: (1) removing the effects of orientation, (2) projecting into the identity subspace [Edwards et al. 1998], (3) projecting across into the subspace of the target model, and (4) adding the appropriate orientation. Figure 26 demonstrates the synthesis of a virtual view of a novel face using this method. Results of tracking a face across large pose variations and predicting novel views were reported on a limited dataset of about 15 short sequences.

Earlier work on multiview-based methods [Beymer 1993] was extended to explore the prior class information that is specific to a face class and can be learned from a set of prototypes [Beymer 1993, 1995]. The key idea of these methods is the vectorized representation of the images at each pose; this is similar to view-based AAM [Cootes et al. 2000]. A vectorized representation at each pose consists of both shape and texture, which are mapped into the standard/average reference shape. The reference shape is computed off-line by averaging shapes consisting of manually defined line segments surrounding the eyes, eyebrows, nose, mouth, and facial outline. The shape-free texture is represented either by the original geometrically normalized prototype images or by PCA bases constructed from these images. Given a new image, a vectorization procedure (similar to the iterative energy minimization procedure in AAM [Cootes et al. 2001]) is invoked that iterates between a *shape step* and a *texture step*. In the texture step, the input image is warped onto a previously computed alignment with the reference shape and then projected into the eigen-subspace. In the shape step, the PCA-reconstructed image is used to compute the alignment for next iteration. In both methods [Beymer 1995; Beymer and Poggio 1995], an optical flow algorithm is used to compute a dense correspondence between the images. To synthesize a virtual view at pose θ_2 of a novel image at pose θ_1, the flow between these poses of the prototype images is computed and then warped to the novel image after the correspondence between the new image and the prototype image at pose θ_1 is computed; using the warped flow, a virtual view can be generated by warping the novel image. Figure 27 illustrates a particular procedure adopted in Beymer and Poggio [1995]: the *parallel deformation* needed to compute the flow between the prototype image and the novel image. An obvious drawback of this approach is the difficulty of computing flow when the prototype image and novel image are dramatically different. To handle this issue, Beymer [1995] proposed first subsampling the estimated dense flow to locate local features (line segments) based on prior knowledge about both images, and then matching the local features. Feeding the virtual views into a simple recognizer based on templates of eyes, nose, and mouth, a recognition rate of 85% was reported on a test set of 620 images (62 people, 10 views per person) given one single real view. Apparently this method is not adequate, since it needs to synthesize all virtual views. A better strategy is to detect the pose of the novel face and synthesize only the prototype (say) frontal view.

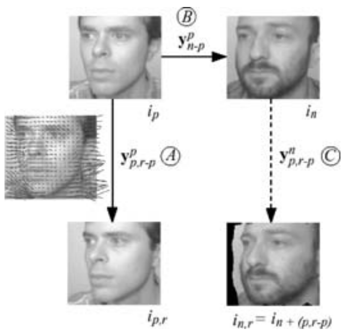

Fig. 27. View synthesis by parallel deformation. First (A) the prototype flow is measured between the prototype image and the novel image at the same pose, then (B) the flow is mapped onto the novel face, and finally (C) the novel face is 2D-warped to the virtual view [Beymer and Poggio 1995].

6.2.3. Single-Image-Based Approaches. Finally, the third class of approaches includes low-level feature-based methods, invariant-feature-based methods, and 3D model-based methods. In Manjunath et al. [1992], a Gabor wavelet-based feature extraction method is proposed for face recognition which is robust to small-angle rotations. In these methods, face shape is usually represented by either a polygonal model or a mesh model which simulates tissue. Due to its complexity and computational cost, no serious attempt to apply this approach to face recognition has been made, except for Gordon [1991], where 3D range data was available. In Zhao and Chellappa [2000b], a unified approach was proposed to solving both the pose and illumination problems. This method is a natural extension of the method proposed in Zhao and Chellappa [2000] to handle the illumination problem. Using a generic 3D model, they approximately solved the correspondence problem involved in a 3D rotation, and performed an input-to-prototype image computation. To address the varying albedo issue in the estimation of both pose and light source, the use of a *self-ratio image* was proposed. The self-ratio image $r_I[x, y]$ was defined as

$$r_I[x, y] = \frac{I[x, y] - I[-x, y]}{I[x, y] + I[-x, y]} = \frac{p[x, y]P_s}{1 + q[x, y]Q_s}, \tag{18}$$

where $I[x, y]$ is the original image and $I[-x, y]$ is the mirrored image.

Using the self-ratio image, which is albedo-free, the authors formulated the following combined estimation problem for pose θ and light source (α, τ):

$$(\theta^*, \alpha^*, \tau^*) = \arg_{\theta,\alpha,\tau} \min[r_{Im}(\alpha, \tau) - r_I(\theta, \alpha, \tau)]^2, \tag{19}$$

where $r_{I(\theta,\alpha,\tau)}$ is the self-ratio image for the virtual frontal view synthesized from the original rotated image I_R via image warping and texture mapping, and r_{Im} is the self-ratio image generated from the 3D face model. Improved recognition results based on subspace LDA [Zhao et al. 1999] were reported on a small database consisting of frontal and quasiprofile images of 115 novel objects (size 48×42). In these experiments, the frontal view images served as the gallery images and nonfrontal view images served as the probe images. Unfortunately, estimation of a single pose value for all the images was done manually. For many images, this estimate was not good, negating the performance improvement.

7. SUMMARY AND CONCLUSIONS

In this paper we have presented an extensive survey of machine recognition of human faces and a brief review of related psychological studies. We have considered two types of face recognition tasks: one from still images and the other from video. We have categorized the methods used for each type, and discussed their characteristics and their pros and cons. In addition to a detailed review of representative work, we have provided summaries of current developments and of challenging issues. We have also identified two important issues in practical face recognition systems: the illumination problem and the pose problem. We have categorized proposed methods of solving these problems and discussed the pros and cons of these methods. To emphasize the importance of system evaluation, three sets of evaluations were described: FERET, FRVT, and XM2VTS.

Getting started in performing experiments in face recognition is very easy. The Colorado State University's Evaluation of Face Recognition Algorithms Web site, http: //www.cs.colostate.edu/evalfacerec/, has an archive of baseline face recognition algorithms. Baseline algorithms available are PCA, LDA, elastic bunch graph matching, and Bayesian Intrapersonal/Extrapersoanl Image Diffference Classifier. Source code, and scripts for running the algorithms can be downloaded. The Web site includes scripts for running the FERET Sep96 evaluation protocol (the FERET data set needs to be obtained from the FERET Web site). The baseline algorithms and FERET Sep96 protocol provide a framework for benchmarking new algorithms. The scripts can be modified to run different sets of images against the baseline. For on-line resources related to face recognition, such as research papers and databases, see Table V.

We give below a concise summary of our discussion, followed by our conclusions, in the same order as the topics have appeared in this paper:

—Machine recognition of faces has emerged as an active research area spanning disciplines such as image processing, pattern recognition, computer vision, and neural networks. There are numerous applications of FRT to commercial systems such as face verification-based ATM and access control, as well as law enforcement applications to video surveillance, etc. Due to its user-friendly nature, face recognition will remain a powerful tool in spite of the existence of very reliable methods of biometric personal identification such as fingerprint analysis and iris scans.

—Extensive research in psychophysics and the neurosciences on human recognition of faces is documented in the literature. We do not feel that machine recognition of faces should strictly follow what is known about human recognition of faces, but it is beneficial for engineers who design face recognition systems to be aware of the relevant findings. On the other hand, machine systems provide tools for conducting studies in psychology and neuroscience.

—Numerous methods have been proposed for face recognition based on image intensities [Chellappa et al. 1995]. Many of these methods have been successfully applied to the task of face recognition, but they have advantages and disadvantages. The choice of a method should be based on the specific requirements of a given task. For example, the EBGM-based method [Okada et al. 1998] has very good performance, but it requires an image size, for example, 128×128, which severely restricts its possible application to video-based surveillance where the image size of the face area is very small. On the other hand, the subspace LDA method [Zhao et al. 1999] works well for both large and small images, for example, 96×84 or 12×11.

—Recognition of faces from a video sequence (especially a surveillance video) is still one of the most challenging problems in face recognition because video is of low quality and the images are small. Often, the subjects of interest are not cooperative, for example, not looking into the camera. One particular difficulty in these applications is how to obtain good-quality gallery images. Nevertheless, video-based face recognition systems using multiple cues have demonstrated good results in relatively controlled environments.

—A crucial step in face recognition is the evaluation and benchmarking of algorithms. Two of the most important face databases and their associated evaluation methods have been reviewed: the FERET, FRVT, and XM2VTS protocols. The availability of these evaluations has had a significant impact on progress in the development of face recognition algorithms.

—Although many face recognition techniques have been proposed and have shown significant promise, robust face recognition is still difficult. There are at least three major challenges: illumination, pose, and recognition in outdoor imagery. A detailed review of methods proposed to solve these problems has been presented. Some basic problems remain to be solved; for example, pose discrimination is not difficult but accurate pose estimation is hard. In addition to these two problems, there are other even more difficult ones, such as recognition of a person from images acquired years apart.

—The impressive face recognition capability of the human perception system has one limitation: the number and types of faces that can be easily distinguished. Machines, on the other hand, can store and potentially recognize as many people as necessary. Is it really possible that a machine can be built that mimics the human perceptual system without its limitations on number and types?

To conclude our paper, we present a conjecture about face recognition based on psychological studies and lessons learned from designing algorithms. We conjecture that different mechanisms are involved in human recognition of familiar and unfamiliar faces. For example, it is possible that 3D head models are constructed, by extensive training for familiar faces, but for unfamiliar faces, multiview 2D images are stored. This implies that we have full probability density functions for familiar faces, while for unfamiliar faces we only have discriminant functions.

REFERENCES

ADINI, Y., MOSES, Y., AND ULLMAN, S. 1997. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 721–732.

AKAMATSU, S., SASAKI, T., FUKAMACHI, H., MASUI, N., AND SUENAGA, Y. 1992. An accurate and robust face identification scheme. In *Proceedings, International Conference on Pattern Recognition*. 217–220.

ATICK, J., GRIFFIN, P., AND REDLICH, N. 1996. Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural Computat. 8*, 1321–1340.

AZARBAYEJANI, A., STARNER, T., HOROWITZ, B., AND PENTLAND, A. 1993. Visually controlled graphics. *IEEE Trans. Patt. Anal. Mach. Intell. 15*, 602–604.

BACHMANN, T. 1991. Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European J. Cog. Psych. 3*, 87–103.

BAILLY-BAILLIERE, E., BENGIO, S., BIMBOT, F., HAMOUZ, M., KITTLER, J., MARIETHOZ, J., MATAS, J., MESSER, K., POPOVICI, V., POREE, F., RUIZ, B., AND THIRAN, J. P. 2003. The BANCA database and evaluation protocol. In *Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication*. 625–638.

BARTLETT, J. C. AND SEARCY, J. 1993. Inversion and configuration of faces. *Cog. Psych. 25*, 281–316.

BARTLETT, M. S., LADES, H. M., AND SEJNOWSKI, T. 1998. Independent component representation for face recognition. In *Proceedings, SPIE Symposium on Electronic Imaging: Science and Technology*. 528–539.

BASRI, R. AND JACOBS, D. W. 2001. Lambertian reflectances and linear subspaces. In *Proceedings, International Conference on Computer Vision*. Vol. II. 383–390.

BELHUMEUR, P. N., HESPANHA, J. P., AND KRIEGMAN, D. J. 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 711–720.

BELHUMEUR, P. N. AND KRIEGMAN, D. J. 1997. What is the set of images of an object under all possible lighting conditions? In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 52–58.

BELL, A. J. AND SEJNOWSKI, T. J. 1995. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation 7*, 1129–1159.

BELL, A. J. AND SEJNOWSKI, T. J. 1997. The independent components of natural scenes are edge filters. *Vis. Res. 37*, 3327–3338.

BEVERIDGE, J. R., SHE, K., DRAPER, B. A., AND GIVENS, G. H. 2001. A nonparametric statisical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. (An updated version can be found online at http://www.cs.colostate.edu/evalfacerec/news.html.)

BEYMER, D. 1995. Vectorizing face images by interleaving shape and texture computations. MIT AI Lab memo 1537. Massachusetts Institute of Technology, Cambridge, MA.

BEYMER, D. J. 1993. Face recognition under varying pose. Tech. Rep. 1461. MIT AI Lab, Massachusetts Institute of Technology, Cambridge, MA.

BEYMER, D. J. AND POGGIO, T. 1995. Face recognition from one example view. In *Proceedings, International Conference on Computer Vision*. 500–507.

BIEDERMAN, I. 1987. Recognition by components: A theory of human image understanding. *Psych. Rev. 94*, 115–147.

BIEDERMAN, I. AND KALOCSAI, P. 1998. Neural and psychophysical analysis of object and face recognition. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 3–25.

BIGUN, J., DUC, B., SMERALDI, F., FISCHER, S., AND MAKAROV, A. 1998. Multi-modal person authentication. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 26–50.

BLACK, M., FLEET, D., AND YACOOB, Y. 1998. A Framework for modelling appearance change in image sequences. In *Proceedings, International Conference on Computer Vision*, 660–667.

BLACK, M. AND YACOOB, Y. 1995. Tracking and recognizing facial expressions in image sequences using local parametrized models of image motion. Tech. rep. CS-TR-3401. Center for Automation Research, Unversity of Maryland, College Park, MD.

BLACKBURN, D., BONE, M., AND PHILLIPS, P. J. 2001. Face recognition vendor test 2000. Tech. rep. http://www.frvt.org.

BLANZ, V. AND VETTER, T. 1999. A Morphable model for the synthesis of 3D faces. In *Proceedings, SIGGRAPH'99*, 187–194.

BLANZ, V. AND VETTER, T. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Patt. Anal. Mach. Intell. 25*, 1063–1074.

BLEDSOE, W. W. 1964. The model method in facial recognition. Tech. rep. PRI:15, Panoramic research Inc., Palo Alto, CA.

BRAND, M. AND BHOTIKA, R. 2001. Flexible flow for 3D nonrigid tracking and shape recovery. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

BRENNAN, S. E. 1985. The caricature generator. *Leonardo, 18*, 170–178.

BRONSTEIN, A., BRONSTEIN, M., GORDON, E., AND KIMMEL, R. 2003. 3D face recognition using geometric invariants. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*.

BRUCE, V. 1988. *Recognizing faces*, Lawrence Erlbaum Associates, London, U.K.

BRUCE, V., BURTON, M., AND DENCH, N. 1994. What's distinctive about a distinctive face? *Quart. J. Exp. Psych. 47A*, 119–141.

BRUCE, V., HANCOCK, P. J. B., AND BURTON, A. M. 1998. Human face perception and identification. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 51–72.

BRUNER, I. S. AND TAGIURI, R. 1954. The perception of people. In *Handbook of Social Psychology*, Vol. 2, G. Lindzey, Ed., Addison-Wesley, Reading, MA, 634–654.

BUHMANN, J., LADES, M., AND MALSBURG, C. V. D. 1990. Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings, International Joint Conference on Neural Networks*. 411–416.

CHELLAPPA, R., WILSON, C. L., AND SIROHEY, S. 1995. Human and machine recognition of faces: A survey. *Proc. IEEE, 83*, 705–740.

CHOUDHURY, T., CLARKSON, B., JEBARA, T., AND PENTLAND, A. 1999. Multimodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*. 176–181.

COOTES, T., TAYLOR, C., COOPER, D., AND GRAHAM, J. 1995. Active shape models—their training and application. *Comput. Vis. Image Understand. 61*, 18–23.

COOTES, T., WALKER, K., AND TAYLOR, C. 2000. View-based active appearance models. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 2001. Active appearance models. *IEEE Trans. Patt. Anal. Mach. Intell. 23*, 681–685.

COX, I. J., GHOSN, J., AND YIANILOS, P. N. 1996. Feature-based face recognition using mixture-distance. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 209–216.

CRAW, I. AND CAMERON, P. 1996. Face recognition by computer. In *Proceedings, British Machine Vision Conference*. 489–507.

DARWIN, C. 1972. *The Expression of the Emotions in Man and Animals*. John Murray, London, U.K.

DECARLO, D. AND METAXAS, D. 2000. Optical flow constraints on deformable models with applications to face tracking. *Int. J. Comput. Vis. 38*, 99–127.

DONATO, G., BARTLETT, M. S., HAGER, J. C., EKMAN, P., AND SEJNOWSKI, T. J. 1999. Classifying facial actions. *IEEE Trans. Patt. Anal. Mach. Intell. 21*, 974–989.

EDWARDS, G. J., TAYLOR, C. J., AND COOTES, T. F. 1998. Learning to identify and track faces in image sequences. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*.

EKMAN, P. Ed., 1998. *Charles Darwin's The Expression of the Emotions in Man and Animals, Third Edition, with Introduction, Afterwords and Commentaries by Paul Ekman*. HarperCollins/Oxford University Press, New York, NY/London, U.K.

ELLIS, H. D. 1986. Introduction to aspects of face processing: Ten questions in need of answers. In *Aspects of Face Processing*, H. Ellis, M. Jeeves, F. Newcombe, and A. Young, Eds. Nijhoff, Dordrecht, The Netherlands, 3–13.

ETEMAD, K. AND CHELLAPPA, R. 1997. Discriminant analysis for recognition of human face images. *J. Opt. Soc. Am. A 14*, 1724–1733.

FISHER, R. A. 1938. The statistical utilization of multiple measuremeents. *Ann. Eugen. 8*, 376–386.

FREEMAN, W. T. AND TENENBAUM, J. B. 2000. Separating style and contents with bilinear models. *Neural Computat. 12*, 1247–1283.

FUKUNAGA, K. 1989. *Statistical Pattern Recognition*, Academic Press, New York, NY.

GALTON, F. 1888. Personal identification and description. *Nature*, (June 21), 173–188.

GAUTHIER, I., BEHRMANN, M., AND TARR, M. J. 1999. Can face recognition really be dissociated from object recognition? *J. Cogn. Neurosci. 11*, 349–370.

GAUTHIER, I. AND LOGOTHETIS, N. K. 2000. Is face recognition so unique after All? *J. Cogn. Neuropsych. 17*, 125–142.

GEORGHIADES, A. S., BELHUMEUR, P. N., AND KRIEGMAN, D. J. 1999. Illumination-based image synthesis: Creating novel images of human faces under differing pose and lighting. In *Proceedings, Workshop on Multi-View Modeling and Analysis of Visual Scenes*, 47–54.

GEORGHIADES, A. S., BELHUMEUR, P. N., AND KRIEGMAN, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Patt. Anal. Mach. Intell. 23*, 643–660.

GEORGHIADES, A. S., KRIEGMAN, D. J., AND BELHUMEUR, P. N. 1998. Illumination cones for recognition under variable lighting: Faces. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 52–58.

GINSBURG, A. G. 1978. Visual information processing based on spatial filters constrained by biological data. AMRL Tech. rep. 78–129.

GONG, S., MCKENNA, S., AND PSARROU, A. 2000. *Dynamic Vision: From Images to Face Recognition*. World Scientific, Singapore.

GORDON, G. 1991. Face recognition based on depth maps and surface curvature. In *SPIE Proceedings, Vol. 1570: Geometric Methods in Computer Vision*. SPIE Press, Bellingham, WA 234–247.

GU, L., LI, S. Z., AND ZHANG, H. J. 2001. Learning probabilistic distribution model for multiview face dectection. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

HAGER, G. D., AND BELHUMEUR, P. N. 1998. Efficient region tracking with parametri models of geometry and illumination. *IEEE Trans. Patt. Anal. Mach. Intell. 20*, 1–15.

HALLINAN, P. W. 1991. Recognizing human eyes. In *SPIE Proceedings, Vol. 1570: Geometric Methods In Computer Vision*. 214–226.

HALLINAN, P. W. 1994. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 995–999.

HANCOCK, P., BRUCE, V., AND BURTON, M. 1998. A comparison of two computer-based face recognition systems with human perceptions of faces. *Vis. Res. 38*, 2277–2288.

HARMON, L. D. 1973. The recognition of faces. *Sci. Am. 229*, 71–82.

HEISELE, B., SERRE, T., PONTIL, M., AND POGGIO, T. 2001. Component-based face detection. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

HILL, H. AND BRUCE, V. 1996. Effects of lighting on matching facial surfaces. *J. Exp. Psych.: Human Percept. Perform. 22*, 986–1004.

HILL, H., SCHYNS, P. G., AND AKAMATSU, S. 1997. Information and viewpoint dependence in face recognition. *Cognition 62*, 201–222.

HJELMAS, E. AND LOW, B. K. 2001. Face detection: A Survey. *Comput. Vis. Image Understand. 83*, 236–274.

HORN, B. K. P. AND BROOKS, M. J. 1989. *Shape from Shading*. MIT Press, Cambridge, MA.

HUANG, J., HEISELE, B., AND BLANZ, V. 2003. Component-based face recognition with 3D morphable models. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*.

ISARD, M. AND BLAKE, A. 1996. Contour tracking by stochastic propagation of conditional density. In *Proceedings, European Conference on Computer Vision*.

JACOBS, D. W., BELHUMEUR, P. N., AND BASRI, R. 1998. Comparing images under variable illumination. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 610–617.

JEBARA, T., RUSSEL, K., AND PENTLAND, A. 1998. Mixture of eigenfeatures for real-time structure from texture. Tech. rep. TR-440, MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA.

JOHNSTON, A., HILL, H., AND CARMAN, N. 1992. Recognizing faces: Effects of lighting direction, inversion and brightness reversal. *Cognition 40*, 1–19.

KALOCSAI, P. K., ZHAO, W., AND ELAGIN, E. 1998. Face similarity space as perceived by humans and artificial systems. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*. 177–180.

KANADE, T. 1973. *Computer recognition of human faces.* Birkhauser, Basel, Switzerland, and Stuttgart, Germany.

KELLY, M. D. 1970. Visual identification of people by computer. Tech. rep. AI-130, Stanford AI Project, Stanford, CA.

KIRBY, M. AND SIROVICH, L. 1990. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Patt. Anal. Mach. Intell. 12*.

KLASEN, L. AND LI, H. 1998. Faceless identification. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 513–527.

KNIGHT, B. AND JOHNSTON, A. 1997. The role of movement in face recognition. *Vis. Cog. 4*, 265–274.

KRUGER, N., POTZSCH, M., AND MALSBURG, C. V. D. 1997. Determination of face position and pose with a learned representation based on labelled graphs. *Image Vis. Comput. 15*, 665–673.

KUNG, S. Y. AND TAUR, J. S. 1995. Decision-based neural networks with signal/image classification applications. *IEEE Trans. Neural Netw. 6*, 170–181.

LADES, M., VORBRUGGEN, J., BUHMANN, J., LANGE, J., MALSBURG, C. V.D., WURTZ, R., AND KONEN, W. 1993. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput. 42*, 300–311.

LANITIS, A., TAYLOR, C. J., AND COOTES, T. F. 1995. Automatic face identification system using flexible appearance models. *Image Vis. Comput. 13*, 393–401.

LAWRENCE, S., GILES, C. L., TSOI, A. C., AND BACK, A. D. 1997. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neural Netw. 8*, 98–113.

LI, B. AND CHELLAPPA, R. 2001. Face verification through tracking facial features. *J. Opt. Soc. Am. 18*.

LI, S. Z. AND LU, J. 1999. Face recognition using the nearest feature line method. *IEEE Trans. Neural Netw. 10*, 439–443.

LI, Y., GONG, S., AND LIDDELL, H. 2001a. Constructing facial identity surfaces in a nonlinear discriminating space. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

LI, Y., GONG, S., AND LIDDELL, H. 2001b. Modelling face dynamics across view and over time. In *Proceedings, International Conference on Computer Vision*.

LIN, S. H., KUNG, S. Y., AND LIN, L. J. 1997. Face recognition/detection by probabilistic decision-based neural network. *IEEE Trans. Neural Netw. 8*, 114–132.

LIU, C. AND WECHSLER, H. 2000a. Evolutionary pursuit and its application to face recognition. *IEEE Trans. Patt. Anal. Mach. Intell. 22*, 570–582.

LIU, C. AND WECHSLER, H. 2000b. Robust coding scheme for indexing and retrieval from large face databases. *IEEE Trans. Image Process. 9*, 132–137.

LIU, C. AND WECHSLER, H. 2001. A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Trans. Image Process. 10*, 598–608.

LIU, J. AND CHEN, R. 1998. Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc. 93*, 1031–1041.

MANJUNATH, B. S., CHELLAPPA, R., AND MALSBURG, C. V. D. 1992. A feature based approach to face recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 373–378.

MARR, D. 1982. *Vision*. W. H. Freeman, San Francisco, CA.

MARTINEZ, A. 2002. Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. *IEEE Trans. Patt. Anal. Mach. Intell. 24*, 748–763.

MARTINEZ, A. AND KAK, A. C. 2001. PCA versus LDA. *IEEE Trans. Patt. Anal. Mach. Intell. 23*, 228–233.

MAURER, T. AND MALSBURG, C. V. D. 1996a. Single-view based recognition of faces rotated in depth. In *Proceedings, International Workshop on Automatic Face and Gesture Recognition*. 176–181.

MAURER, T. AND MALSBURG, C. V. D. 1996b. Tracking and learning graphs and pose on image sequences of faces. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*. 176–181.

MCKENNA, S. J. AND GONG, S. 1997. Non-intrusive person authentication for access control by visual tracking and face recognition. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*. 177–183.

MCKENNA, S. AND GONG, S. 1998. Recognising moving faces. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 578–588.

MATAS, J. ET AL., 2000. Comparison of face verification results on the XM2VTS database. In *Proceedings, International Conference on Pattern Recognition*, Vol. 4, 858–863.

MESSER, K., MATAS, J., KITTLER, J., LUETTIN, J., AND MAITRE, G. 1999. XM2VTSDB: The Extended M2VTS Database. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*. 72–77.

MIKA, S., RATSCH, G., WESTON, J., SCHOLKOPF, B., AND MULLER, K.-R. 1999. Fisher discriminant analysis with kernels. In *Proceedings, IEEE Workshop on Neural Networks for Signal Processing*.

MOGHADDAM, B., NASTAR, C., AND PENTLAND, A. 1996. A Bayesian similarity measure for direct image matching. In *Proceedings, International Conference on Pattern Recognition*.

MOGHADDAM, B. AND PENTLAND, A. 1997. Probabilistic visual learning for object representation. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 696–710.

MOON, H. AND PHILLIPS, P. J. 2001. Computational and performance aspects of PCA-based face recognition algorithms. *Perception, 30*, 301–321.

MURASE, H. AND NAYAR, S. 1995. Visual learning and recognition of 3D objects from appearances. *Int. J. Comput. Vis. 14*, 5–25.

NEFIAN, A. V. AND HAYES III, M. H. 1998. Hidden Markov models for face recognition. In *Proceedings, International Conference on Acoustics, Speech and Signal Processing.* 2721–2724.

OKADA, K., STEFFANS, J., MAURER, T., HONG, H., ELAGIN, E., NEVEN, H., AND MALSBURG, C. V. D. 1998. The Bochum/USC Face Recognition System and how it fared in the FERET Phase III Test. In *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 186–205.

O'TOOLE, A. J., ROARK, D., AND ABDI, H. 2002. Recognitizing moving faces. A psychological and neural synthesis. *Trends Cogn. Sci. 6*, 261–266.

PANTIC, M. AND ROTHKRANTZ, L. J. M. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Patt. Anal. Mach. Intell. 22*, 1424–1446.

PENEV, P. AND SIROVICH, L. 2000. The global dimensionality of face space. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*.

PENEV, P. AND ATICK, J. 1996. Local feature analysis: A general statistical theory for objecct representation. *Netw.: Computat. Neural Syst. 7*, 477–500.

PENTLAND, A., MOGHADDAM, B., AND STARNER, T. 1994. View-based and modular eigenspaces for face recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

PERKINS, D. 1975. A definition of caricature and recognition. *Stud. Anthro. Vis. Commun. 2*, 1–24.

PHILLIPS, P. J., GROTHER, P. J., MICHEALS, R. J., BLACKBURN, D. M., TABASSI, E., AND BONE, J. M. 2003. Face recognition vendor test 2002: Evaluation report. NISTIR 6965, 2003. Available online at http://www.frvt.org.

PHILLIPS, P. J. 1998. Support vector machines applied to face fecognition. *Adv. Neural Inform. Process. Syst. 11,* 803–809.

PHILLIPS, P. J., MCCABE, R. M., AND CHELLAPPA, R. 1998. Biometric image processing and recognition. In *Proceedings, European Signal Processing Conference*.

PHILLIPS, P. J., MOON, H., RIZVI, S., AND RAUSS, P. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Patt. Anal. Mach. Intell. 22*.

PHILLIPS, P. J., WECHSLER, H., HUANG, J., AND RAUSS, P. 1998b. The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput. 16*, 295–306.

PIGEON, S. AND VANDENDORPE, L. 1999. The M2VTS multimodal face database (Release 1.00). In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*. 403–409.

RIKLIN-RAVIV, T. AND SHASHUA, A. 1999. The quotient image: Class based re-rendering and recognition with varying illuminations. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 566–571.

RIZVI, S. A., PHILLIPS, P. J., AND MOON, H. 1998. A verification protocol and statistical performance analysis for face recognition algorithms. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 833–838.

ROWLEY, H. A., BALUJA, S., AND KANADE, T. 1998. Neural network based face detection. *IEEE Trans. Patt. Anal. Mach. Intell. 20*.

CHOUDHURY, A. K. R. AND CHELLAPPA, R. 2003. Face reconstruction from monocular video using uncertainty analysis and a generic model. *Comput. Vis. Image Understand. 91*, 188–213.

RUDERMAN, D. L. 1994. The statistics of natural images. *Netw.: Comput. Neural Syst. 5*, 598–605.

SALI, E. AND ULLMAN, S. 1998. Recognizing novel 3-D objects under new illumination and viewing position using a small number of example views or even a single view. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 153–161.

SAMAL, A. AND IYENGAR, P. 1992. Automatic recognition and analysis of human faces and facial expressions: A survey. *Patt. Recog. 25*, 65–77.

SAMARIA, F. 1994. Face recognition using hidden markov models. Ph.D. dissertation. University of Cambridge, Cambridge, U.K.

SAMARIA, F. AND YOUNG, S. 1994. HMM based architecture for face identification. *Image Vis. Comput. 12*, 537–583.

SCHNEIDERMAN, H. AND KANADE, T. 2000. Probabilistic modelling of local Appearance and spatial reationships for object recognition. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*. 746–751.

SERGENT, J. 1986. Microgenesis of face perception. In *Aspects of Face Processing*, H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, Eds. Nijhoff, Dordrecht, The Netherlands.

SHASHUA, A. 1994. Geometry and photometry in 3D visual recognition. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, MA.

SHEPHERD, J. W., DAVIES, G. M., AND ELLIS, H. D. 1981. Studies of cue saliency. In *Perceiving and Remembering Faces*, G. M. Davies, H. D. Ellis, and J. W. Shepherd, Eds. Academic Press, London, U.K.

SHIO, A. AND SKLANSKY, J. 1991. Segmentation of people in motion. In *Proceedings, IEEE Workshop on Visual Motion*. 325–332.

SIROVICH, L. AND KIRBY, M. 1987. Low-dimensional procedure for the characterization of human face. *J. Opt. Soc. Am. 4*, 519–524.

STEFFENS, J., ELAGIN, E., AND NEVEN, H. 1998. PersonSpotter—fast and robust system for human detection, tracking and recognition. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*. 516–521.

STROM, J., JEBARA, T., BASU, S., AND PENTLAND, A. 1999. Real time tracking and modeling of faces: An EKF-based analysis by synthesis approach. Tech. rep. TR-506, MIT Media Lab, Massachusetts, Institute of Technology, Cambridge, MA.

SUNG, K. AND POGGIO, T. 1997. Example-based learning for view-based human face detection. *IEEE Trans. Patt. Anal. Mach. Intell. 20*, 39–51.

SWETS, D. L. AND WENG, J. 1996b. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Patt. Anal. Mach. Intell. 18*, 831–836.

SWETS, D. L. AND WENG, J. 1996. Discriminant analysis and eigenspace partition tree for face and object recognition from views. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*. 192–197.

TARR, M. J. AND BULTHOFF, H. H. 1995. Is human object recognition better described by geon structural descriptions or by multiple views—comment on Biederman and Gerhardstein (1993). *J. Exp. Psych.: Hum. Percep. Perf. 21*, 71–86.

TERZOPOULOS, D. AND WATERS, K. 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Patt. Anal. Mach. Intell. 15*, 569–579.

THOMPSON, P. 1980. Margaret Thatcher—A new illusion. *Perception, 9*, 483–484.

TSAI, P. S. AND SHAH, M. 1994. Shape from shading using linear approximation. *Image Vis. Comput. 12*, 487–498.

TRIGGS, B., MCLAUCHLAN, P., HARTLEY, R., AND FITZGIBBON, A. 2000. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, Springer-Verlag, Berlin, Germany.

TURK, M. AND PENTLAND, A. 1991. Eigenfaces for recognition. *J. Cogn. Neurosci. 3*, 72–86.

ULLMAN, S. AND BASRI, R. 1991. Recognition by linear combinations of models. *IEEE Trans. Patt. Anal. Mach. Intell. 13*, 992–1006.

VAPNIK, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

VETTER, T. AND POGGIO, T. 1997. Linear object classes and image synthesis from a single example image. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 733–742.

VIOLA, P. AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*.

WECHSLER, H., KAKKAD, V., HUANG, J., GUTTA, S., AND CHEN, V. 1997. Automatic video-based person authentication using the RBF network. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*. 85–92.

WILDER, J. 1994. Face recognition using transform coding of gray scale projection and the neural tree network. In *Artificial Neural Networks with Applications in Speech and Vision*, R. J. Mammone, Ed. Chapman Hall, New York, NY, 520–536.

WISKOTT, L., FELLOUS, J.-M., AND VON DER MALSBURG, C. 1997. Face recognition by elastic bunch graph matching. *IEEE Trans. Patt. Anal. Mach. Intell. 19*, 775–779.

YANG, M. H., KRIEGMAN, D., AND AHUJA, N. 2002. Detecting faces in images: A survey. *IEEE Trans. Patt. Anal. Mach. Intell. 24*, 34–58.

YIN, R. K. 1969. Looking at upside-down faces. *J. Exp, Psych. 81*, 141–151.

YUILLE, A. L., COHEN, D. S., AND HALLINAN, P. W. 1992. Feature extractiong from faces using deformable templates. *Int. J. Comput. Vis. 8*, 99–112.

YUILLE, A. AND HALLINAN, P. 1992. Deformable templates. In *Active vision*, A. Blake, and A. Yuille, Eds., Cambridge, MA, 21–38.

ZHAO, W. 1999. *Robust Image Based 3D Face Recognition*, Ph.D. dissertation. University of Maryland, College Park, MD.

ZHAO, W. AND CHELLAPPA, R. 2000b. SFS Based View synthesis for robust face recognition. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*.

ZHAO, W. AND CHELLAPPA, R. 2000. Illumination-insensitive face recognition using symmetric shape-from-shading. In *Proceedings, Conference on Computer Vision and Pattern Recognition*. 286–293.

ZHAO, W., CHELLAPPA, R., AND KRISHNASWAMY, A. 1998. Discriminant analysis of principal components for face recognition. In *Proceedings, International Conference on Automatic Face and Gesture Recognition*. 336–341.

ZHAO, W., CHELLAPPA, R., AND PHILLIPS, P. J. 1999. Subspace linear discriminant analysis for face recognition. Tech. rep. CAR-TR-914, Center for Automation Research, University of Maryland, College Park, MD.

ZHOU, S., KRUEGER, V., AND CHELLAPPA, R. 2003. Probabilistic recognition of human faces from video. *Comput. Vis. Image Understand. 91*, 214–245.

Received July 2002; accepted June 2003