Chapter 1
Introduction

Stan Z. Li and Anil K. Jain

1.1 Face Recognition

Face recognition is a task that humans perform routinely and effortlessly in our daily lives. Wide availability of powerful and low-cost desktop and embedded computing systems has created an enormous interest in automatic processing of digital images in a variety of applications, including biometric authentication, surveillance, human-computer interaction, and multimedia management. Research and development in automatic face recognition follows naturally.

Face recognition has several advantages over other biometric modalities such as fingerprint and iris: besides being natural and nonintrusive, the most important advantage of face is that it can be captured at a distance and in a covert manner. Among the six biometric attributes considered by Hietmeyer [16], facial features scored the highest compatibility in a Machine Readable Travel Documents (MRTD) [27] system based on a number of evaluation factors, such as enrollment, renewal, machine requirements, and public perception, shown in Fig. 1.1. Face recognition, as one of the major biometric technologies, has become increasingly important owing to rapid advances in image capture devices (surveillance cameras, camera in mobile phones), availability of huge amounts of face images on the Web, and increased demands for higher security.

The first automated face recognition system was developed by Takeo Kanade in his Ph.D. thesis work [18] in 1973. There was a dormant period in automatic face recognition until the work by Sirovich and Kirby [19, 38] on a low dimen-

S.Z. Li (✉)
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: szli@cbsr.ia.ac.cn

A.K. Jain
Michigan State University, East Lansing, MI 48824, USA
e-mail: jain@cse.msu.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_1, © Springer-Verlag London Limited 2011

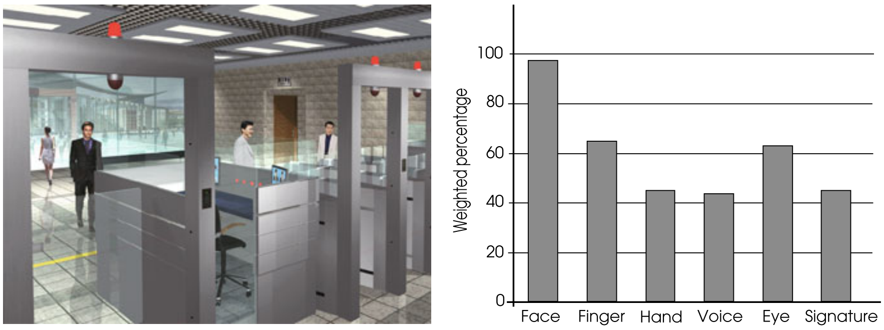


Fig. 1.1 A scenario of using biometric MRTD systems for passport control (*left*), and a comparison of various biometric traits based on MRTD compatibility (*right*, from Hietmeyer [16] with permission)

sional face representation, derived using the Karhunen–Loeve transform or Principal Component Analysis (PCA). It is the pioneering work of Turk and Pentland on Eigenface [42] that reinvigorated face recognition research. Other major milestones in face recognition include: the Fisherface method [3, 12], which applied Linear Discriminant Analysis (LDA) after a PCA step to achieve higher accuracy; the use of local filters such as Gabor jets [21, 45] to provide more effective facial features; and the design of the AdaBoost learning based cascade classifier architecture for real time face detection [44].

Face recognition technology is now significantly advanced since the time when the Eigenface method was proposed. In the constrained situations, for example where lighting, pose, stand-off, facial wear, and facial expression can be controlled, automated face recognition can surpass human recognition performance, especially when the database (gallery) contains a large number of faces.[1] However, automatic face recognition still faces many challenges when face images are acquired under unconstrained environments. In the following sections, we give a brief overview of the face recognition process, analyze technical challenges, propose possible solutions, and describe state-of-the-art performance.

This chapter provides an introduction to face recognition research. Main steps of face recognition processing are described. Face detection and recognition problems are explained from a face subspace viewpoint. Technology challenges are identified and possible strategies for solving some of the problems are suggested.

1.2 Categorization

As a biometric system, a face recognition system operates in either or both of two modes: (1) face verification (or authentication), and (2) face identification (or recognition). Face verification involves a one-to-one match that compares a query face

[1] Most individuals can identify only a few thousand people in real life.

Part I
Face Image Modeling and Representation

Chapter 2
Face Recognition in Subspaces

Gregory Shakhnarovich and Baback Moghaddam

2.1 Introduction

Images of faces, represented as high-dimensional pixel arrays, often belong to a manifold of intrinsically low dimension. Face recognition, and computer vision research in general, has witnessed a growing interest in techniques that capitalize on this observation and apply algebraic and statistical tools for extraction and analysis of the underlying manifold. In this chapter, we describe in roughly chronologic order techniques that identify, parameterize, and analyze linear and nonlinear subspaces, from the original Eigenfaces technique to the recently introduced Bayesian method for probabilistic similarity analysis. We also discuss comparative experimental evaluation of some of these techniques as well as practical issues related to the application of subspace methods for varying pose, illumination, and expression.

2.2 Face Space and Its Dimensionality

Computer analysis of face images deals with a visual signal (light reflected off the surface of a face) that is registered by a digital sensor as an array of pixel values. The pixels may encode color or only intensity. In this chapter, we assume the latter case (i.e., gray-level imagery). After proper normalization and resizing to a fixed m-by-n size, the pixel array can be represented as a point (i.e., vector) in an mn-dimensional *image space* by simply writing its pixel values in a fixed (typically raster) order. A critical issue in the analysis of such multidimensional data is the

G. Shakhnarovich (✉)
Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA
e-mail: gregory@ai.mit.edu

B. Moghaddam
Mitsubishi Electric Research Labs, Cambridge, MA 02139, USA
e-mail: baback@merl.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_2, © Springer-Verlag London Limited 2011

Chapter 3
Face Subspace Learning

Wei Bian and Dacheng Tao

3.1 Introduction

The last few decades have witnessed a great success of subspace learning for face recognition. From principal component analysis (PCA) [43] and Fisher's linear discriminant analysis [1], a dozen of dimension reduction algorithms have been developed to select effective subspaces for the representation and discrimination of face images [17, 21, 45, 46, 51]. It has demonstrated that human faces, although usually represented by thousands of pixels encoded in high-dimensional arrays, they are intrinsically embedded in a vary low dimensional subspace [37]. The using of subspace for face representation helps to reduce "the curse of dimensionality" in subsequent classification, and suppress variations of lighting conditions and facial expressions. In this chapter, we first briefly review conventional dimension reduction algorithms and then present the trend of recent dimension reduction algorithms for face recognition.

The earliest subspace method for face recognition is Eigenface [43], which uses PCA [23] to select the most representative subspace for representing a set of face images. It extracts the principal eigenspace associated with a set of training face images. Mathematically, PCA maximizes the variance in the projected subspace for a given dimensionality, decorrelates the training face images in the projected subspace, and maximizes the mutual information between appearance (training face images) and identity (the corresponding labels) by assuming that face images are Gaussian distributed. Thus, it has been successfully applied for face recognition. By projecting face images onto the subspace spanned by Eigenface, classifiers can be used in the subspace for recognition. One main limitation of Eigenface is that the

W. Bian (✉) · D. Tao
Centre for Quantum Computation & Intelligence Systems, FEIT, University of Technology, Sydney, NSW 2007, Australia
e-mail: wei.bian@student.uts.edu.au

D. Tao
e-mail: dacheng.tao@uts.edu.au

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_3, © Springer-Verlag London Limited 2011

class labels of face images cannot be explored in the process of learning the projection matrix for dimension reduction. Another representative subspace method for face recognition is Fisherface [1]. In contrast to Eigenface, Fisherface finds class specific linear subspace. The dimension reduction algorithm used in Fisherface is Fisher's linear discriminant analysis (FLDA), which simultaneously maximizes the between-class scatter and minimizes the within-class scatter of the face data. FLDA finds in the feature space a low dimensional subspace where the different classes of samples remain well separated after projection to this subspace. If classes are sampled from Gaussian distributions, all with identical covariance matrices, then FLDA maximizes the mean value of the KL divergences between different classes. In general, Fisherface outperforms Eigenface due to the utilized discriminative information.

Although FLDA shows promising performance on face recognition, it has the following major limitations. FLDA discards the discriminative information preserved in covariance matrices of different classes. FLDA models each class by a single Gaussian distribution, so it cannot find a proper projection for subsequent classification when samples are sampled from complex distributions, for example, mixtures of Gaussians. In face recognition, face images are generally captured with different expressions or poses, under different lighting conditions and at different resolution, so it is more proper to assume face images from one person are mixtures of Gaussians. FLDA tends to merge classes which are close together in the original feature space. Furthermore, when the size of the training set is smaller than the dimension of the feature space, FLDA has the undersampled problem.

To solve the aforementioned problems in FLDA, a dozen of variants have been developed in recent years. Especially, the well-known undersample problem of FLDA has received intensive attention. Representative algorithms include the optimization criterion for generalized discriminant analysis [44], the unified subspace selection framework [44] and the two stage approach via QR decomposition [52]. Another important issue is that FLDA meets the class separation problem [39]. That is because FLDA puts equal weights on all class pairs, although intuitively close class pairs should contribute more to the recognition error [39]. To reduce this problem, Lotlikar and Kothari [30] developed the fractional-step FLDA (FS-FLDA) by introducing a weighting function. Loog et al. [28] developed another weighting method for FLDA, namely the approximate pairwise accuracy criterion (aPAC). The advantage of aPAC is that the projection matrix can be obtained by the eigenvalue decomposition. Both methods use weighting schemes to select a subspace that better separates close class pairs. Recently, the general mean [39] (including geometric mean [39] and harmonic mean [3]) base subspace selection and the max-min distance analysis (MMDA) [5] have been proposed to adaptively choose the weights.

Manifold learning is a new technique for reducing the dimensionality in face recognition and has received considerable attentions in recent years. That is because face images lie in a low-dimensional manifold. A large number of algorithms have been proposed to approximate the intrinsic manifold structure of a set of face images, such as locally linear embedding (LLE) [34], ISOMAP [40], Laplacian eigenmaps (LE) [2], Hessian eigenmaps (HLLE) [11], Generative Topographic Mapping

(GTM) [6] and local tangent space alignment (LTSA) [53]. LLE uses linear coefficients, which reconstruct a given measurement by its neighbors, to represent the local geometry, and then seeks a low-dimensional embedding, in which these coefficients are still suitable for reconstruction. ISOMAP preserves global geodesic distances of all pairs of measurements. LE preserves proximity relationships by manipulations on an undirected weighted graph, which indicates neighbor relations of pairwise measurements. LTSA exploits the local tangent information as a representation of the local geometry and this local tangent information is then aligned to provide a global coordinate. Hessian Eigenmaps (HLLE) obtains the final low-dimensional representations by applying eigen-analysis to a matrix which is built by estimating the Hessian over neighborhood. All these algorithms have the out of sample problem and thus a dozen of linearizations have been proposed, for example, locality preserving projections (LPP) [20] and discriminative locality alignment (DLA) [55]. Recently, we provide a systematic framework, that is, patch alignment [55], for understanding the common properties and intrinsic difference in different algorithms including their linearizations. In particular, this framework reveals that: i) algorithms are intrinsically different in the patch optimization stage; and ii) all algorithms share an almost-identical whole alignment stage. Another unified view of popular manifold learning algorithms is the graph embedding framework [48]. It is shown that manifold learning algorithms are more effective than conventional dimension reduction algorithms, for example, PCA and FLDA, in exploiting local geometry information.

In contrast to conventional dimension reduction algorithms that obtain a low dimensional subspace with each basis being a linear combination of all the original high dimensional features, sparse dimension reduction algorithms [9, 24, 59] select bases composed by only a small number of features of the high dimensional space. The sparse subspace is more interpretable both psychologically and physiologically. One popular sparse dimension reduction algorithm is sparse PCA, which generalizes the standard PCA by imposing sparsity constraint on the basis of the low dimensional subspace. The Manifold elastic net (MEN) [56] proposed recently is another sparse dimension reduction algorithm. It obtains a sparse projection matrix by imposing the elastic net penalty (i.e., the combination of the lasso penalty and the L_2-norm penalty) over the loss (i.e., the criterion) of a discriminative manifold learning, and formulates the problem as lasso which can be efficiently solved. In sum, sparse learning has many advantages, because (1) sparsity can make the data more succinct and simpler, so the calculation of the low dimensional representation and the subsequent recognition becomes more efficient. Parsimony is especially important for large scale face recognition systems; (2) sparsity can control the weights of original variables and decrease the variance brought by possible over-fitting with the least increment of the bias. Therefore, the learn model can generalize better and obtain high recognition rate for distorted face images; and (3) sparsity provides a good interpretation of a model, thus reveals an explicit relationship between the objective of the model and the given variables. This is important for understanding face recognition.

One fundamental assumption in face recognition, including dimension reduction, is that the training and test samples are independent and identically distributed (i.i.d.) [22, 31, 38]. It is, however, very possible that this assumption does not hold, for example, the training and test face images are captured under different expressions, postures or lighting conditions, letting alone test subjects do not even appear in the training set [38]. Transfer learning has emerged as a new learning scheme to deal with such problem. By properly utilizing the knowledge obtained from the auxiliary domain task (training samples), it is possible to boost the performance on the target domain task (test samples). The idea of cross domain knowledge transfer was also introduced to subspace learning [31, 38]. It has shown that by using transfer subspace learning, the recognition performance on the cases where the face images in training and test sets are not identically distributed can be significantly improved compared with comparison against conventional subspace learning algorithms.

The rest of this chapter presents three groups of dimension reduction algorithms for face recognition. Specifically, Sect. 3.2 presents the general mean criterion and the max-min distance analysis (MMDA). Section 3.3 is dedicated to manifold learning algorithms, including the discriminative locality alignment (DLA) and manifold elastic net (MEN). The transfer subspace learning framework is presented in Sect. 3.4. In all of these sections, we first present principles of algorithms and then show thorough empirical studies.

3.2 Subspace Learning—A Global Perspective

Fisher's linear discriminant analysis (FLDA) is one of the most well-known methods for linear subspace selection, and has shown great value in subspace based face recognition. Being developed by Fisher [14] for binary-class classification and then generalized by Rao [33] for multiple-class tasks, FLDA utilizes the ratio of the between-class to within-class scatter as a definition of discrimination. It can be verified that under the homoscedastic Gaussian assumption, FLDA is Bayes optimal [18] in selecting a $c-1$ dimensional subspace, wherein c is the class number. Suppose there are c classes, represented by homoscedastic Gaussians $N(\mu_i, \varSigma \mid \omega_i)$ with the prior probability p_i, $1 \le i \le c$, where μ_i is the mean of class ω_i and $\varSigma$ is the common covariance. The Fisher's criterion is given by [15]

$$\max_W \operatorname{tr}\bigl(\bigl(W^{\mathrm{T}} \varSigma W\bigr)^{-1} W^{\mathrm{T}} S_b W\bigr) \tag{3.1}$$

where

$$S_b = \sum_{i=1}^{c} p_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}, \quad \text{with } \mu = \sum_{i=1}^{c} p_i \mu_i. \tag{3.2}$$

It has been pointed out that the Fisher's criterion implies the maximization of the arithmetic mean of the pairwise distances between classes in the subspace. To see this, let us first define the distance between classes ω_i and ω_j in the subspace W as

$$\varDelta(\omega_i, \omega_i \mid W) = \operatorname{tr}\bigl(\bigl(W^{\mathrm{T}} \varSigma W\bigr)^{-1} W^{\mathrm{T}} D_{ij} W\bigr), \quad \text{with } D_{ij} = (\mu_i - \mu_j)(\mu_i - \mu_j)^{\mathrm{T}}. \tag{3.3}$$

Chapter 4
Local Representation of Facial Features

Joni-Kristian Kämäräinen, Abdenour Hadid, and Matti Pietikäinen

The aim of this chapter is to give a comprehensive overview of different facial representations and in particular describe local facial features.

4.1 Introduction

Developing face recognition systems involves two crucial issues: facial representation and classifier design [47, 101]. The aim of facial representation is to derive a set of features from the raw face images which minimizes the intra-class variations (i.e., within face instances of a same individual) and maximizes the extra-class variations (i.e., between face images of different individuals). Obviously, if inadequate facial representations are adopted, even the most sophisticated classifiers fail to accomplish the face recognition task. Therefore, it is important to carefully decide on what facial representation to adopt when designing face recognition systems. Ideally, the facial feature representation should: (i) discriminate different individuals well while tolerating within-class variations; (ii) be easily extracted from the raw face images in order to allow fast processing; and (iii) lie in a low dimensional space (short vector length) in order to avoid a computationally expensive classifier. Naturally, it is

J.-K. Kämäräinen (✉)
Machine Vision and Pattern Recognition Laboratory, Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: Joni.Kamarainen@lut.fi

A. Hadid · M. Pietikäinen
Machine Vision Group, Dept. of Electrical and Information Engineering, University of Oulu, Oulu, Finland

A. Hadid
e-mail: hadid@ee.oulu.fi

M. Pietikäinen
e-mail: mkp@ee.oulu.fi

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_4, © Springer-Verlag London Limited 2011

not easy to find features which meet all these criteria because of the large variability in facial appearances due to different imaging factors such as scale, orientation, pose, facial expressions, lighting conditions, aging, presence of glasses, etc. These considerations are important for the other subtasks in face biometrics: detection, localization and registration, and verification, and thus, a key issue in face recognition is finding efficient facial feature representations.

Numerous methods have been proposed in literature for representing facial images for recognition purposes. The earliest attempts, such as Kanade's work in early 70s [41], are based on representing faces in terms of geometrical relationships, such as distances and angles, between the facial landmarks (eyes, mouth etc.). Later, appearance based techniques have been proposed. These methods generally consider a face as a 2D array of pixels and aim at deriving descriptors for face appearance without explicit use of face geometry. Following these lines, different holistic methods such as Principal Component Analysis (PCA) [82], Linear Discriminant Analysis (LDA) [21] and the more recent 2D PCA [92] have been widely studied. Lately local descriptors have gained an increasing attention due to their robustness to challenges such as pose and illumination changes. Among these descriptors are Gabor filters and Local Binary Patterns [2] which are shown to be very successful in encoding facial appearance.

4.1.1 Structure and Scope of the Chapter

The aim of this chapter is to give a comprehensive overview of different facial representations and in particular describe local facial features. Section 4.2 discusses the major methods which have been proposed in literature. Then, more detailed descriptions of two widely used approaches, namely local binary patterns and Gabor filters, are presented in Sects. 4.3 and 4.4, respectively. Section 4.5 discusses related issues and promising directions. Finally, concluding remarks are drawn in Sect. 4.6.

The methods discussed in this chapter can be applied to detection and recognition of faces or face parts (landmarks). Face parts are also referred to as facial features, but we use the terms feature and facial feature interchangeably for any features extracted from the face area. We specifically discuss local binary patterns in the context of face recognition and Gabor features in the context of face part detection, but they can be used in the both tasks. Furthermore, the feature extraction methods are discussed from the face image processing point of view and other face description methods are available for the modeling purposes, such as the active shape models and morphable model described in the following chapters. These novel modeling methods can also be applied to face recognition without explicit feature extraction and classification as discussed in this chapter.

Chapter 5
Face Alignment Models

Phil Tresadern, Tim Cootes, Chris Taylor, and Vladimir Petrović

5.1 Introduction

In building models of facial appearance, we adopt a statistical approach that learns the ways in which the shape and texture of the face vary across a range of images. We rely on obtaining a suitably large and representative training set of images of faces, each of which is annotated with a set of feature points that define correspondences across the set. The positions of the feature points also define the shape of the face, and are analysed to learn the ways in which the shape can vary. The patterns of intensities are analysed in a similar way to learn how the texture can vary. The result is a model which is capable of synthesising any of the training images and generalising from them, but is specific enough that only face-like images are generated.

To build a statistical appearance model, we require a set of training images that covers the types of variation we want the model to represent. For instance, if we are only interested in faces with neutral expressions, we need only include neutral expressions in the model. If, however, we want to synthesise and recognise a range of expressions, the training set should include images of people smiling, frowning, winking and so on. Ideally, the faces in the training set should be of at least as high a resolution as those in the images we wish to synthesise or interpret.

5.1.1 Statistical Models of Shape

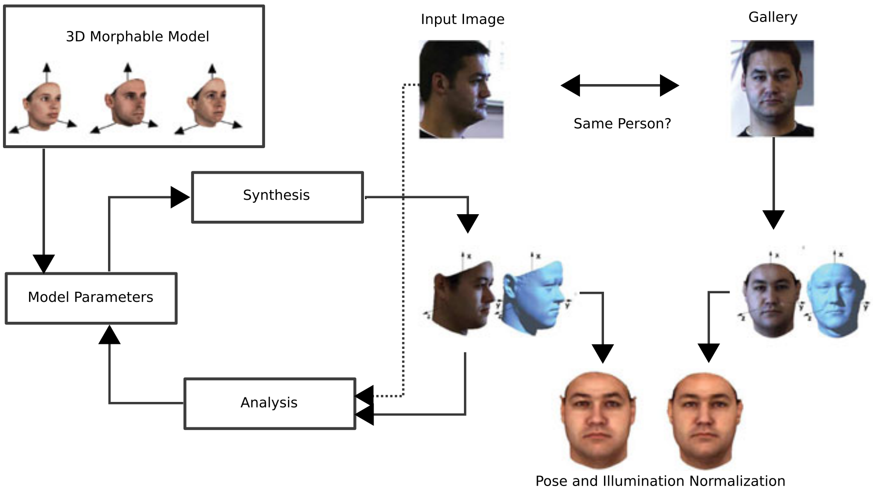
To define a shape model, we first annotate each face with a fixed number of points that define the key facial features (and their correspondences across the training set) and represent the shape of the face in the image. Typically, we place points around

P. Tresadern · T. Cootes (✉) · C. Taylor · V. Petrović
Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK
e-mail: t.cootes@man.ac.uk

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_5, © Springer-Verlag London Limited 2011

Chapter 6
Morphable Models of Faces

Reinhard Knothe, Brian Amberg, Sami Romdhani, Volker Blanz, and Thomas Vetter



R. Knothe (✉) · B. Amberg · S. Romdhani · T. Vetter
Department of Mathematics and Computer Science, University of Basel, Bernoullistrasse 16, 4056 Basel, Switzerland
e-mail: reinhard.knothe@unibas.ch

B. Amberg
e-mail: brian.amberg@unibas.ch

S. Romdhani
e-mail: sami.romdhani@unibas.ch

T. Vetter
e-mail: thomas.vetter@unibas.ch

V. Blanz
Universität Siegen, Hölderlinstrasse 3, 57068 Siegen, Germany
e-mail: blanz@mpi-sb.mpg.de

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_6, © Springer-Verlag London Limited 2011

6.1 Introduction

Our approach is based on an *analysis by synthesis* framework. In this framework, an input image is analyzed by searching for the parameters of a generative model such that the generated image is as similar as possible to the input image. The parameters are then used for high-level tasks such as identification.

To be applicable to all input face images, a good model must be able to generate all possible face images. Face images vary widely with respect to the imaging conditions (illumination and the position of the camera relative to the face, called pose) and with respect to the identity and the expression of the face. A generative model must not only allow for these variations but must also separate the sources of variation such that e.g. the identity can be determined regardless of pose, illumination or expression.

In this chapter, we present the Morphable Model, a three-dimensional (3D) representation that enables the accurate modeling of any illumination and pose as well as the separation of these variations from the rest (identity and expression). The Morphable Model is a generative model consisting of a linear 3D shape and appearance model plus an imaging model, which maps the 3D surface onto an image. The 3D shape and appearance are modeled by taking linear combinations of a training set of example faces. We show that linear combinations yield a realistic face only if the set of example faces is in correspondence. A good generative model should accurately distinguish faces from nonfaces. This is encoded in the probability distribution over the model parameters, which assigns a high probability to faces and a low probability to nonfaces. The distribution is learned together with the shape and appearance space from the training data.

Based on these principles, we detail the construction of a 3D Morphable Face Model in Sect. 6.2. The main step of model construction is to build the correspondences of a set of 3D face scans. Such models have become a well-established technology which is able to perform various tasks, not only face recognition, but also face image analysis [6] (e.g., estimating the 3D shape from a single photograph), expression transfer from one photograph to another [10, 46], animation of faces [10], training of feature detectors [22, 24], and stimuli generation for psychological experiments [29] to name a few. The power of these models comes at the cost of an expensive and tedious construction process, which has led the scientific community often to focus on more easily constructed but less powerful models. Recently, a complete 3D Morphable Face Model built from 3D face scans, the *Basel Face Model* (BFM), was made available to the public (faces.cs.unibas.ch) [34]. An alternative approach to construct a 3D Morphable Model is to generate the model directly from a video sequence [12] using nonrigid structure from motion. While this requires far less manual intervention, it also results in a less detailed and inaccurate model.

With a good generative face model, we are half the way to a face recognition system. The remaining part of the system is the face analysis algorithm (the *fitting algorithm*). The fitting algorithm finds the parameters of the model that generate an image which is as close as possible to the input image. In this chapter, we focus on fitting the model to a single image. We detail two fitting algorithms in Sect. 6.4.

Based on these two fitting algorithms, identification results are presented in Sect. 6.5 for face images varying in illumination and pose, as well as for 3D face scans. The fitting methods presented here are energy minimization methods, a different class of fitting methods are regression based, and try to learn a correspondence between appearance and model coefficients. For 2D Models, [20] proposed to learn a linear regression mapping the residual between a current estimate and the final fit, and [27] proposed to use a support vector regression on Haar features of the image to directly predict 6 coefficients of a 2D mouth model from 12 Haar features of mouth images.

6.1.1 Three-Dimensional Representation

Each individual face can generate a variety of images when seen from different viewpoints, under different illumination and with different expressions. This huge diversity of face images makes their analysis difficult. In addition to the general differences between individual faces, the appearance variations in images of a single faces can be separated into the following four sources.

- Pose changes can result in dramatic changes in images. Due to self-occlusions different parts of the object become visible or invisible. Additionally, the parts seen in two views change their spatial configuration relative to each other.
- Illumination changes influence the appearance of a face even if the pose of the face is fixed. The distribution of light sources around a face changes the brightness distribution in the image, the locations of attached shadows, and specular reflections. Additionally, cast shadows can generate prominent contours in facial images.
- Facial expressions are another source of variations in images. Only a few facial landmarks that are directly coupled with the bony structure of the skull, such as the corners of the eye or the position of the earlobes, are constant in a face. Most other features can change their spatial configuration or position via articulation of the jaw or muscle action (e.g., moving eyebrows, lips, or cheeks).
- On a longer timescale faces change because of aging, change of hairstyle, and the use of makeup or accessories.

The isolation and explicit description of these sources of variations must be the ultimate goal of a face analysis system. For example, it is desirable that the parameters that code the identity of a person are not perturbed by a modification of pose. In an analysis by synthesis framework, this implies that the face model must account for each of these variations independently by explicit parameters.

We need a generative model which is a concise description of the observed phenomena. The image of a face is generated according to the laws of physics which describe the interaction of light with the face surface and the camera. The parameters for pose and illumination can therefore be described most concisely when modeling the face as a 3D surface. A concise description of the variability of human faces on the other hand can not be derived from physics. We therefore describe the variations in 3D shape and albedo of human faces with parameters learned from examples.

Chapter 7
Illumination Modeling for Face Recognition

Ronen Basri and David Jacobs

7.1 Introduction

Changes in lighting can produce large variability in the appearance of faces, as illustrated in Fig. 7.1. Characterizing this variability is fundamental to understanding how to account for the effects of lighting on face recognition. In this chapter, we will discuss solutions to a problem: Given (1) a three-dimensional description of a face, its pose, and its reflectance properties, and (2) a 2D query image, how can we efficiently determine whether lighting conditions exist that can cause this model to produce the query image? We describe methods that solve this problem by producing simple, linear representations of the set of all images a face can produce under all lighting conditions. These results can be directly used in face recognition systems that capture 3D models of all individuals to be recognized. They also have the potential to be used in recognition systems that compare strictly 2D images but that do so using generic knowledge of 3D face shapes.

One way to measure the difficulties presented by lighting, or any variability, is the number of degrees of freedom needed to describe it. For example, the pose of a face relative to the camera has six degrees of freedom—three rotations and three translations. Facial expression has a few tens of degrees of freedom if one considers the number of muscles that may contract to change expression. To describe the light that strikes a face, we must describe the intensity of light hitting each point on the face from each direction. That is, light is a function of position and direction, meaning that light has an infinite number of degrees of freedom. In this chapter, however, we will show that effective systems can account for the effects of lighting

Portions of this chapter are reprinted, with permission, from Basri and Jacobs [6], © 2003 IEEE.

R. Basri (✉)
The Weizmann Institute of Science, Rehovot 76100, Israel
e-mail: ronen.basri@weizmann.ac.il

D. Jacobs
University of Maryland, College Park, MD 20742, USA
e-mail: djacobs@umiacs.umd.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_7, © Springer-Verlag London Limited 2011

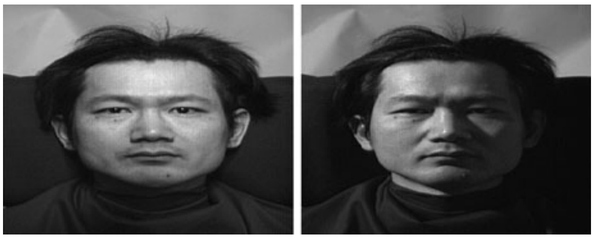


Fig. 7.1 Same face under different lighting conditions

using fewer than 10 degrees of freedom. This can have considerable impact on the speed and accuracy of recognition systems.

Support for low-dimensional models is both empirical and theoretical. Principal component analysis (PCA) on images of a face obtained under various lighting conditions shows that this image set is well approximated by a low-dimensional, linear subspace of the space of all images (see, e.g., [19]). Experimentation shows that algorithms that take advantage of this observation can achieve high performance, for example, [17, 21].

In addition, we describe theoretical results that, with some simplified assumptions, prove the validity of low-dimensional, linear approximations to the set of images produced by a face. For these results, we assume that light sources are distant from the face, but we do allow arbitrary combinations of point sources (e.g., the Sun) and diffuse sources (e.g., the sky). We also consider only diffuse components of reflectance, modeled as Lambertian reflectance, and we ignore the effects of cast shadows, such as those produced by the nose. We do, however, model the effects of attached shadows, as when one side of a head faces away from a light. Theoretical predictions from these models provide a good fit to empirical observations and produce useful recognition systems. This suggests that the approximations made capture the most significant effects of lighting on facial appearance. Theoretical models are valuable not only because they provide insight into the role of lighting in face recognition, but also because they lead to analytically derived, low-dimensional, linear representations of the effects of lighting on facial appearance, which in turn can lead to more efficient algorithms.

An alternate stream of work attempts to compensate for lighting effects without the use of 3D face models. This work directly matches 2D images using representations of images that are found to be insensitive to lighting variations. These include image gradients [12], Gabor jets [29], the direction of image gradients [13, 24], and projections to subspaces derived from linear discriminants [8]. A large number of these methods are surveyed in [50]. These methods are certainly of interest, especially for applications in which 3D face models are not available. However, methods based on 3D models may be more powerful, as they have the potential to compensate completely for lighting changes, whereas 2D methods cannot achieve such invariance [1, 13, 35]. Another approach of interest, the Morphable Model, is to use general 3D knowledge of faces to improve methods of image comparison.

Chapter 8
Face Recognition Across Pose and Illumination

Ralph Gross, Simon Baker, Iain Matthews, and Takeo Kanade

8.1 Introduction

The most recent evaluation of commercial face recognition systems shows the level of performance for face verification of the best systems to be on par with fingerprint recognizers for frontal, uniformly illuminated faces [38]. Recognizing faces reliably across changes in pose and illumination has proved to be a much more difficult problem [9, 24, 38]. Although most research has so far focused on frontal face recognition, there is a sizable body of work on pose invariant face recognition and illumination invariant face recognition. However, face recognition across pose *and* illumination has received little attention.

8.1.1 Multiview Face Recognition and Face Recognition Across Pose

Approaches addressing pose variation can be classified into two categories depending on the type of gallery images they use. Multiview face recognition is a direct extension of frontal face recognition in which the algorithms require gallery images of every subject at every pose. In face recognition across pose, we are concerned

R. Gross (✉) · S. Baker · I. Matthews · T. Kanade
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: rgross@cs.cmu.edu

S. Baker
e-mail: simonb@cs.cmu.edu

I. Matthews
e-mail: iainm@cs.cmu.edu

T. Kanade
e-mail: tk@cs.cmu.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_8, © Springer-Verlag London Limited 2011

Chapter 9
Skin Color in Face Analysis

J. Birgitta Martinkauppi, Abdenour Hadid, and Matti Pietikäinen

9.1 Introduction

Color is a common feature used in machine vision applications. As a cue, it offers several advantages: easy to understand and use. Implementations can be made computationally fast and efficient, thus providing a low level cue. Under stable and uniform illumination, color cue remains robust against geometrical changes. Its ability to separate the targets from background depends on the color dissimilarity between targets and background. In some scenes, the color itself is enough for object detection.

The main difficulty in using color in machine vision applications is that the cameras are not able to distinguish changes of surface colors from color shifts caused by varying illumination spectra. Thus, color is sensitive to changes in illumination which are common under uncontrolled environments. The changes can be due to varying light level, for example, shadowing, varying light color due to changes in spectral power distribution (like daylight and fluorescent light source), or both. Cameras and their settings may produce different appearances which are different from the perception of human vision system.

Several strategies have been employed to reduce the illumination sensitivity. In one strategy, the color information is separated into two components, color inten-

J.B. Martinkauppi (✉)
Department of Electrical Engineering and Automation, University of Vaasa, Wolffintie 34, 65101 Vaasa, Finland
e-mail: birmar@uwasa.fi

A. Hadid · M. Pietikäinen
Machine Vision Group, Department of Electrical and Information Engineering, University of Oulu, P.O. Box 4500, 90014 Oulu, Finland

A. Hadid
e-mail: hadid@ee.oulu.fi

M. Pietikäinen
e-mail: mkp@ee.oulu.fi

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_9, © Springer-Verlag London Limited 2011

sity and color chromaticity. Use of color chromaticity component reduces the effect of varying light levels. To cancel the effect of illumination color and thus different spectral power distributions, numerous color constancy algorithms have been suggested, but their success has been limited [6]. A different strategy to these is to tolerate or adapt the model to the illumination changes. This strategy can produce promising results even under drastic variations in target colors as shown in this chapter for facial recognition.

It is often preferable to get rid as much as possible of the dependencies on lighting intensity. The perfect case would be to also cancel-out the effect of the illuminant color (by defining a color representation which is only a function of the surface reflectance) but, thus far this has not been achieved in machine vision. The human visual system is superior in this sense, since human visual perception in which the color is perceived by the eye depends quite significantly on surface reflectance, although the light reaching the eye is a function of surface reflectance, illuminant color and lighting intensity.

For face detection, color has been an intriguing and popular cue. It is often used as a preprocessing step to select regions of interests for further, more computationally demanding processing. For instance, with the appearance-based face detection, an exhaustive scan (at different locations and scales) of the images is conducted when searching for the faces [54]. However, when the color cue is available, one can reduce the search regions by pre-processing the images and selecting the skin-like areas only.

This chapter deals with the role of color in facial image analysis such as face detection and recognition. First, we introduce the use of color information in the field of facial image analysis in particular (Sect. 9.2). Then, in Sect. 9.3, we give an introduction to color formation and discuss the effect of illumination on color appearance, and its consequences. The skin data can come from different sources like real faces, photos or print. Separating the sources of skin data is presented in Sect. 9.4, and skin color modeling is discussed in Sect. 9.5. Section 9.6 reviews the use of color in face detection, while the contribution of color to face recognition is covered in Sect. 9.7. Finally, conclusions are drawn in Sect. 9.8.

9.2 Color Cue and Facial Image Analysis

The properties of the face pattern pose a very difficult problem for facial image analysis: a face is a dynamic and nonrigid object which is difficult to handle. Its appearance varies due to changes in pose, expressions, illuminations and other factors such as age and make-up. As a consequence, most of the facial analysis tasks generally involve heavy computations due to the complexity of facial patterns. Therefore, one may need some additional cues, such as color or motion, in order to assist and accelerate the analysis. These additional cues also offer an indication of the reliability of the face analysis results: the more the cues support the analysis, the more one can be confident about the results. For instance, with the appearance-based face

Chapter 10
Face Aging Modeling

Unsang Park and Anil K. Jain

10.1 Introduction

Face recognition accuracy is typically limited by the large intra-class variations caused by factors such as pose, lighting, expression, and age [16]. Therefore, most of the current work on face recognition is focused on compensating for the variations that degrade face recognition performance. However, facial aging has not received adequate attention compared to other sources of variations such as pose, lighting, and expression.

Facial aging is a complex process that affects both the shape and texture (e.g., skin tone or wrinkles) of a face. The aging process appears in different manifestations in different age groups, gender and ethnicity. While facial aging is mostly represented by the facial growth in younger age groups (e.g., below 18 years of age), it is mostly represented by relatively large texture changes and minor shape changes (e.g., due to the change of weight or stiffness of skin) in older age groups (e.g., over 18 years of age). Therefore, an age invariant face recognition scheme needs to be able to compensate for both types of aging process.

Some of the face recognition applications where age invariance or correction is required include (i) identifying missing children, (ii) screening for watch list, and (iii) multiple enrollment detection problems. These three scenarios have two common characteristics: (i) a significant age difference exists between probe and gallery images (images obtained at verification and enrollment stages, respectively) and (ii) an inability to obtain a user's face image to update the template (gallery). Identifying missing children is one of the most apparent applications where age compensation is needed to improve the recognition performance. In screening applications, aging is a major source of difficulty in identifying suspects in a watch list. Repeat

U. Park (✉) · A.K. Jain
Michigan State University, East Lansing, MI 48824, USA
e-mail: parkunsa@cse.msu.edu

A.K. Jain
e-mail: jain@cse.msu.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_10, © Springer-Verlag London Limited 2011

offenders commit crimes at different time periods in their lives, often starting as a juvenile and continuing throughout their lives. It is not unusual to encounter a time lapse of ten to twenty years between the first (enrollment) and subsequent (verification) arrests. Multiple enrollment detection for issuing government documents such as driver licenses and passports is a major problem that various government and law enforcement agencies face in the facial databases that they maintain. Face or some other types of biometric traits (e.g., fingerprint or iris) are the only ways to reliably detect multiple enrollments.

Ling et al. [10] studied how age differences affect the face recognition performance in a real passport photo verification task. Their results show that the aging process does increase the recognition difficulty, but it does not surpass the challenges posed due to change in illumination or expression. Studies on face verification across age progression [19] have shown that: (i) simulation of shape and texture variations caused by aging is a challenging task, as factors like life style and environment also contribute to facial changes in addition to biological factors, (ii) the aging effects can be best understood using 3D scans of human head, and (iii) the available databases to study facial aging are not only small but also contain uncontrolled external and internal variations (e.g., pose, illumination, expression, and occlusion). It is due to these reasons that the effect of aging in facial recognition has not been as extensively investigated as other factors that lead to large intra-class variations in facial appearance.

Some biological and cognitive studies on face aging process have also been conducted, see [18, 25]. These studies have shown that cardioidal strain is a major factor in the aging of facial outlines. Such results have also been used in psychological studies, for example, by introducing aging as caricatures generated by controlling 3D model parameters [12]. Patterson et al. [15] compared automatic aging simulation results with forensic sketches and showed that further studies in aging are needed to improve face recognition techniques. A few seminal studies [20, 24] have demonstrated the feasibility of improving face recognition accuracy by simulated aging. There has also been some work done in the related area of age estimation using statistical models, for example, [8, 9]. Geng et al. [7] learn a subspace of aging pattern based on the assumption that similar faces age in similar ways. Their face representation is composed of face texture and the 2D shape represented by the coordinates of the feature points as in the Active Appearance Models. Computer graphics community has also shown facial aging modeling methods in 3D domain [22], but the effectiveness of the aging model was not evaluated by conducting a face recognition test.

Table 10.1 gives a brief comparison of various methods for modeling aging proposed in the literature. The performance of these models is evaluated in terms of the improvement in the identification accuracy. When multiple accuracies were reported in any of the studies under the same experimental setup (e.g., due to different choice of probe and gallery), their average value is listed in Table 10.1; when multiple accuracies are reported under different approaches, the best performance is reported. The identification accuracies of various studies in Table 10.1 cannot be directly compared due to the differences in the database, the number of subjects

Table 10.1 A comparison of various face aging models [13]

	Approach	Face matcher	Database (#subjects, #images) in probe and gallery	Rank-1 identification accuracy (%)	
				Original image	After aging model
Ramanathan et al. (2006) [20]	Shape growth modeling up to age 18	PCA	Private database (109, 109)	8.0	15.0
Lanitis et al. (2002) [8]	Build an aging function in terms of PCA coefficients of shape and texture	Mahalanobis distance, PCA	Private database (12, 85)	57.0	68.5
Geng et al. (2007) [7]	Learn aging pattern on concatenated PCA coefficients of shape and texture across a series of ages	Mahalanobis distance, PCA	FG-NET[*] (10, 10)	14.4	38.1
Wang et al. (2006) [26]	Build an aging function in terms of PCA coefficients of shape and texture	PCA	Private database (NA, 2000)	52.0	63.0
Patterson et al. (2006) [14]	Build an aging function in terms of PCA coefficients of shape and texture	PCA	MORPH[+] (9, 36)	11.0	33.0
Park et al. [13]	Learn aging pattern based on PCA coefficients in separated 3D shape and texture given 2D database	FaceVACS	FG-NET[**] (82, 82)	26.4	37.4
			MORPH-Album1[++] (612,612)	57.8	66.4
			BROWNS (4, 4)—probe (100, 100)—gallery	15.6	28.1

[*] Used only a very small subset of the FG-NET database that contains a total of 82 subjects

[+] Used only a very small subset of the MORPH database that contains a total of 625 subjects

[**] Used all the subjects in FG-NET

[++] Used all the subjects in MORPH-Album1 which have multiple images

and the underlying face recognition method used for evaluation. Usually, the larger the number of subjects and the larger the database variations in terms of age, pose, lighting and expression, the smaller the recognition performance improvement by an aging model. The identification accuracy for each approach in Table 10.1 before aging simulation indicates the difficulty of the experimental setup for the face recognition test as well as the limitations of the face matcher.

Compared with other published approaches, the aging model proposed by Park et al. [13] has the following features.

- 3D aging modeling: Includes a pose correction stage and a more realistic model of the aging pattern in the *3D domain*. Considering that the aging is a 3D process, 3D modeling is better suited to capture the aging patterns. Their method is the only viable alternative to building a 3D aging model directly, as no 3D aging database is currently available. Scanned 3D face data rather than reconstructed is used in [22], but they were not collected for aging modeling and hence, do not contain as much aging information as the 2D facial aging database.
- Separate modeling of shape and texture changes: Three different modeling methods, namely, shape modeling only, separate shape and texture modeling and combined shape and texture modeling (e.g., applying 2nd level PCA to remove the correlation between shape and texture after concatenating the two types of feature vectors) were compared. It has been shown that the separate modeling is better than combined modeling method, given the FG-NET database as the training data.
- Evaluation using a state-of-the-art commercial face matcher, FaceVACS: A state-of-the-art face matcher, FaceVACS from Cognitec [4] has been used to evaluate the aging model. Their method can thus be useful in practical applications requiring an age correction process. Even though their method has been evaluated only on one particular face matcher, it can be used directly in conjunction with any other 2D face matcher.
- Diverse Databases: FG-NET has been used for aging modeling and the aging model has been evaluated on three different databases: FG-NET (in a leave-one-person-out fashion), MORPH, and BROWNS. Substantial performance improvements have been observed on all three databases.

The rest of this Chapter is organized as follows: Sect. 10.2 introduces the preprocessing step of converting 2D images to 3D models, Sect. 10.3 describes the aging model, Sect. 10.4 presents the aging simulation methods using the aging model, and Sect. 10.5 provides experimental results and discussions. Section 10.6 summarizes the conclusions and lists some directions for future work.

10.2 Preprocessing

Park et el. propose to use a set of 3D face images to learn the model for recognition, because the true craniofacial aging model [18] can be appropriately formulated only in 3D. However, since only 2D aging databases are available, it is necessary to first convert these 2D face images into 3D. Major notations that are used in the following sections are defined first.

- $\mathbf{S}_{mm} = \{S_{mm,1}, S_{mm,2}, \ldots, S_{mm,n_{mm}}\}$: a set of 3D face models used in constructing the reduced morphable model. n_{mm} is the number of 3D face models.
- $\mathbf{S}_{\alpha}$: reduced morphable model represented with the model parameter α.
- $\mathbf{S}_{2d,i}^{j} = \{x_1, y_1, \ldots, x_{n_{2d}}, y_{n_{2d}}\}$: 2D facial feature points for the ith subject at age j. n_{2d} is the number of points in 2D.

Chapter 11
Face Detection

Stan Z. Li and Jianxin Wu

11.1 Introduction

Face detection is the first step in automated face recognition. Its reliability has a major influence on the performance and usability of the entire face recognition system. Given a single image or a video, an ideal face detector should be able to identify and locate all the present faces regardless of their position, scale, orientation, age, and expression. Furthermore, the detection should be done irrespectively of extraneous illumination conditions and the image and video content.

Face detection can be performed based on several cues: skin color (for faces in color images and videos), motion (for faces in videos), facial/head shape, facial appearance, or a combination of these parameters. Most successful face detection algorithms are appearance-based without using other cues. The processing is done as follows: An input image is scanned at all possible locations and scales by a subwindow. Face detection is posed as classifying the pattern in the subwindow as either face or nonface. The face/nonface classifier is learned from face and nonface training examples using statistical learning methods.

This chapter presents appearance-based and learning-based methods.[1] It will highlight AdaBoost-based methods because so far they are the most successful ones in terms of detection accuracy and speed. Effective postprocessing methods are also described. Experimental results are provided.

[1] The reader is referred to a review article [50] for other earlier face detection methods.

S.Z. Li (✉)
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: szli@cbsr.ia.ac.cn

J. Wu
School of Computer Engineering, Nanyang Technological University, Singapore, Singapore
e-mail: jxwu@ntu.edu.sg

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_11, © Springer-Verlag London Limited 2011

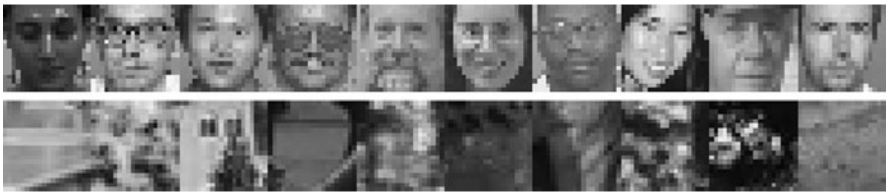


Fig. 11.1 Face (*top*) and nonface (*bottom*) examples

11.2 Appearance and Learning-Based Approaches

With appearance-based methods, face detection is treated as a problem of classifying each scanned subwindow as one of two classes (that is, face and nonface). Appearance-based methods avoid difficulties in modeling 3D structures of faces by considering possible face appearances under various conditions. A face/nonface classifier may be learned from a training set composed of face examples taken under possible conditions as would be seen in the running stage and nonface examples as well (see Fig. 11.1 for a random sample of 10 face and 10 nonface subwindow images). Building such a classifier is possible because pixels on a face are highly correlated, whereas those in a nonface subwindow present much less regularity.

However, large variations brought about by changes in facial appearance, lighting, and expression make the face manifold or face/nonface boundaries highly complex [4, 37, 40]. Changes in facial view (head pose) further complicate the situation. A nonlinear classifier is needed to deal with the complicated situation. The speed is also an important issue for realtime performance.

Great research effort has been made for constructing complex yet fast classifiers and much progress has been achieved since 1990s. Turk and Pentland [41] describe a detection system based on principal component analysis (PCA) subspace or eigenface representation. Whereas only likelihood in the PCA subspace is considered in the basic PCA method, Moghaddam and Pentland [23] also consider the likelihood in the orthogonal complement subspace; using that system, the likelihood in the image space (the union of the two subspaces) is modeled as the product of the two likelihood estimates, which provide a more accurate likelihood estimate for the detection. Sung and Poggio [38] first partition the image space into several face and nonface clusters and then further decompose each cluster into the PCA and null subspaces. The Bayesian estimation is then applied to obtain useful statistical features. The system of Rowley et al.'s [31] uses retinally connected neural networks. Through a sliding window, the input image is examined after going through an extensive preprocessing stage. Osuna et al. [24] train a nonlinear support vector machine to classify face and nonface patterns, and Yang et al. [51] use the SNoW (Sparse Network of Winnows) learning architecture for face detection. In these systems, a bootstrap algorithm is used iteratively to collect meaningful nonface examples from images that do not contain any faces for retraining the detector. Schneiderman and Kanade [34] use multiresolution information for different levels of wavelet transform. A nonlinear classifier is constructed using statistics of products of histograms computed from face and nonface examples. The system of five

Chapter 12
Facial Landmark Localization

Xiaoqing Ding and Liting Wang

12.1 Introduction

Face detection and recognition is a vibrant area of biometrics with active research and commercial efforts over the last 20 years. The task of face detection is to search faces in images, reporting their positions by a bounding box. Recent studies [19, 31] have shown that face detection has already been a state-of-the-art technology in both accuracy and speed. However, face detection is not sufficient to acquire facial landmarks, for example, eye contours, mouth corners, nose, eyebrows, etc. This is therefore the task of facial landmark localization which aims to find the accurate positions of the facial feature points as illustrated in Fig. 12.1. It is a fundamental and significant work in face-related areas, for example, face recognition, face cartoon/sketch, face pose estimate, model-based face tracking, eye/mouth motion analysis, 3D face reconstruction, etc.

There is a wide variety of works related to facial landmark localization. The early researches extract facial landmarks without a global model. Facial landmarks, such as the eye corners and centers, the mouth corners and center, the nose corners, chin and cheek borders are located based on geometrical knowledge. The first step consists of the establishment of a rectangular search region for the mouth and a rectangular search region for the eyes. The borders are extracted by applying corner detection algorithm such as SUSAN border extraction algorithm [17]. Such methods are fast, however, they could not deal with faces of large variation in appearance due to pose, rotation, illumination and background changes.

X. Ding (✉) · L. Wang
State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
e-mail: dingxq@tsinghua.edu.cn

L. Wang
e-mail: wangltmail@tsinghua.edu.cn

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_12, © Springer-Verlag London Limited 2011

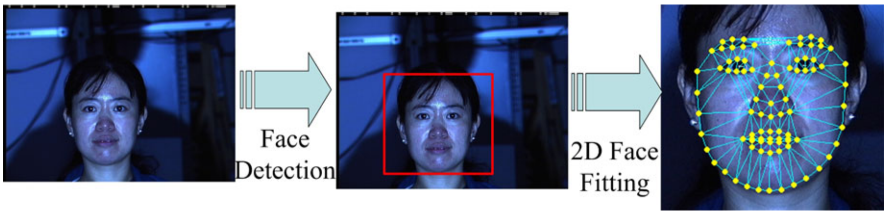


Fig. 12.1 Facial landmark localization

Different with the earlier model-independent algorithm, some researches focus on model-dependent algorithm. Hsu and Jain [18] propose an approach which represents human faces semantically via facial components such as eyes, mouth, face outline, and the hair outline. Each facial component is encoded by a closed (or open) snake that is drawn from a 3D generic face model. The face shape model here is not based on statistical learning and it still could not deal with faces of large variation in appearance due to pose, rotation, illumination and background changes.

With the prominent successful research of Active Shape Model (ASM) [7, 9, 3, 8] and Active Appearance Model (AAM) [4–6, 10–13], face shape is well modeled as a linear combination of principal modes (major eigenvectors) learned from the training face shapes. By learning statistical distribution of shapes and textures from training database, a deformable shape model is built. The boundary of objects with similar shapes to those in the training set could be extracted by fitting this deformable model to images. Depending on the different tasks, ASM and AAM can be built in different ways. On one hand, we might construct a person specific ASM or AAM across pose, illumination, and expression. Such a person-specific model might be useful for interactive user interface applications including head pose estimation, gaze estimation etc. On the other hand, we might construct ASM or AAM to fit any face, including faces unseen in training set. Evidence suggests that the performance of the person-specific facial landmark localization is substantially better than the performance of generic facial landmark localization. As indicated in [15], Gross's experimental results confirm that generic facial landmark localization is far harder than person-specific facial landmark localization and the performance degrades quickly when fitting to images which are unseen in the training set.

In recent years, there are several improved research works based on the framework of AAM. Papandreou and Maragos [27] introduce two enhancements to inverse-compositional AAM matching algorithms in order to overcome the limitation when inverse-compositional AAM matching algorithms are used in conjunction with models exhibiting significant appearance variation, such as AAMs trained on multiple-subject human face images. Liu Xiaoming [25, 26] proposes a discriminative framework to greatly improve the robustness, accuracy and efficiency of face alignment for unseen data. Liebelt et al. [24] develop an iterative multi-level algorithm that combines AAM fitting and robust 3D shape alignment. Xiao et al. [33] also develop the research work of combining 2D AAM and 3D Morphable Model (3DMM). Hamsici and Martinez [16] derive a new approach carries the advantages of AAM and 3DMM that can model nonlinear changes in examples without the need of a pre-alignment step. Lee and Kim [21] propose a tensor-based AAM that can handle a variety of subjects, poses, expressions, and illuminations in the tensor algebra framework. They reported Tensor-based AAM reduced the fitting error of the conventional AAM by about two pixels and the computation time by about 0.6 second.

There are also several improved research works based on the framework of ASM. Tu et al. [30] propose a hierarchical CONDENSATION framework to estimate the face configuration parameter under the framework of ASM. Jiao et al. [20] present a W-ASM, in which Gabor wavelet features are used for modeling local image structure. Zhang and Ai [34] propose an Adaboost discriminative framework which improves the accuracy, efficiency, and robustness of ASM. The same research works are also carried on by Li and Ito [23] who describe a modeling method by using AdaBoosted histogram classifiers. Brunet et al. [2] define a new criterion to select landmarks that have good generalization properties. Vogler et al. [32] combine the ASM with 3D deformable model which governs the overall shape, orientation and location.

In the following, we will introduce a coarse-to-fine facial landmark localization algorithm which uses discriminant learning to remedy the generalization problems based on the framework of Active Shape Model.

12.2 Framework for Landmark Localization

This facial landmark localization framework consists of training and locating procedures, as illustrated in Fig. 12.2.

The training procedure is building a face deformable model via shape modeling and local appearance modeling. This procedure needs a great amount of hand labeled data. The locating procedure consists of firstly the face detection, the eye localization and then the facial landmark localization based on the face deformable model. In the eye localization procedure, we will introduce a robust and precise eye localization method, and then adopt this method to precisely locate the eye position. The eye localization method is real-time. In the facial landmark localization procedure, a random forest embedded active shape model is adopted. In the following paragraphs, they will be presented and discussed in detail.

12.3 Eye Localization

The eye localization is a crucial step towards automatic face recognition and facial landmark localization due to the fact that these face related applications need to normalize faces, measure the relative positions or extract features according to eye positions. Like other problems of object detection under complex scene such as face detection, car detection, eye patterns also have large variation in appearance due to various factors, such as size, pose, rotation, the closure and opening of eyes,

Chapter 13
Face Tracking and Recognition in Video

Rama Chellappa, Ming Du, Pavan Turaga, and Shaohua Kevin Zhou

13.1 Introduction

Faces are expressive three dimensional objects. Information useful for recognition tasks can be found both in the geometry and texture of the face and also facial motion. While geometry and texture together determine the 'appearance' of the face, motion encodes behavioral cues such as idiosyncratic head movements and gestures which can potentially aid in recognition tasks. Traditional face recognition systems have relied on a gallery of still images for learning and a probe of still images for recognition. While the advantage of using motion information in face videos has been widely recognized, computational models for video based face recognition have only recently gained attention.

In this chapter, we consider applications where one is presented with a video sequence—either in a single camera setting or a multi-camera setting—and the goal is to recognize the person in the video. The gallery could consist of either still-images or could be videos themselves.

Video is a rich source of information in that it can lead to potentially better representations by offering more views of the face. Further, the role of facial motion for face perception has been well documented. Psychophysical studies [26] have

R. Chellappa (✉) · M. Du · P. Turaga
Department of Electrical and Computer Engineering, Center for Automation Research, University of Maryland, College Park, MD 20742, USA
e-mail: rama@umiacs.umd.edu

M. Du
e-mail: mingdu@umiacs.umd.edu

P. Turaga
e-mail: pturaga@umiacs.umd.edu

S.K. Zhou
Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA
e-mail: kzhou@scr.siemens.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_13, © Springer-Verlag London Limited 2011

found evidence that when both structure and dynamics information is available, humans tend to rely more on dynamics under nonoptimal viewing conditions (such as low spatial resolution, harsh illumination conditions etc.). Dynamics also aids in recognition of familiar faces [31]. If one were to ignore temporal dependencies, a video sequence can be considered as a collection of still images; so still-image-based recognition algorithms can always be applied. The properties of video sequences that can be exploited are (1) temporal correlations, (2) idiosyncratic dynamic information, and (3) availability of multiple views. Video thus proves useful in various tasks—it can be used to generate better appearance models, mitigate effects of non-cooperative viewing conditions, localize a face using motion, model facial behavior for improved recognition, generate better models of face shape from multiple views, etc.

The rest of the chapter is organized as follows. In Sect. 13.2, we describe the utility of videos in enhancing performance of image-based recognition tasks. In Sect. 13.3, we discuss a joint tracking-recognition framework that allows for using the motion information in a video to better localize and identify the person in the video using still galleries. In Sect. 13.4, we discuss how to jointly capture facial appearance and dynamics to obtain a parametric representation for video-to-video recognition. In Sect. 13.5, we discuss recognition in multi-camera networks where the probe and gallery both consist of multi-camera videos. Finally in Sect. 13.6, we present concluding remarks and directions for future research.

13.2 Utility of Video

Frame-Based Fusion An immediate possible utilization of temporal information for video-based face recognition is to fuse the results obtained by a 2D face recognition algorithm on each frame of the sequence. The video sequence can be seen as an unordered set of images to be used for both training and testing phases. During testing one can use the sequence as a set of probes, each of them providing a decision regarding the identity of the person. Appropriate fusion techniques can then be applied to provide the final identity. Perhaps the most frequently used fusion strategy in this case is majority voting [24, 34].

In [28], Park et al. adopt three matchers for frame-level face recognition: FaceVACS, PCA and correlation. They use the sum rule (with min-max normalization) to fuse results obtained from the three matchers and the maximum rule to fuse results of individual frames. In [21], the concept of identity surface is proposed to represent the hyper-surface formed by projecting face patterns of an individual to the feature vector space parameterized with respect to pose. This surface is learned from gallery videos. In testing stage, model trajectories are synthesized on the identity surfaces of enrolled subjects after the pose parameters of probe video have been estimated. Every point on the trajectory corresponds to a frame of the video and trajectory distance is defined as a weighted sum of point-wise distances. The model trajectory that yields minimum distance to the probe video's trajectory gives the final identification result. Based on the result that images live approximately in a bilinear

Chapter 14
Face Recognition at a Distance

Frederick W. Wheeler, Xiaoming Liu, and Peter H. Tu

14.1 Introduction

Face recognition, and biometric recognition in general, have made great advances in the past decade. Still, the vast majority of practical biometric recognition applications involve cooperative subjects at close range. Face Recognition at a Distance (FRAD) has grown out of the desire to automatically recognize people out in the open, and without their direct cooperation. The face is the most viable biometric for recognition at a distance. It is both openly visible and readily imaged from a distance. For security or covert applications, facial imaging can be achieved without the knowledge of the subject. There is great interest in iris at a distance, however it is doubtful that iris will outperform face with comparable system complexity and cost. Gait information can also be acquired over large distances, but face will likely continue to be a more discriminating identifier.

In this chapter, we will review the primary driving applications for FRAD and the challenges still faced. We will discuss potential solutions to these challenges and review relevant research literature. Finally, we will present a few specific activities to advance FRAD capabilities and discuss expected future trends. For the most part, we will focus our attention on issues that are unique to FRAD. Some of the main challenges of FRAD are shared by many other face recognition applications, and are thoroughly covered in other dedicated chapters of this book.

Distance itself is not really the fundamental motivating factor for FRAD. The real motivation is to work over large coverage areas without subject cooperation.

F.W. Wheeler (✉) · X. Liu · P.H. Tu
Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY 12309, USA
e-mail: wheeler@ge.com

X. Liu
e-mail: liux@ge.com

P.H. Tu
e-mail: tu@ge.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_14, © Springer-Verlag London Limited 2011

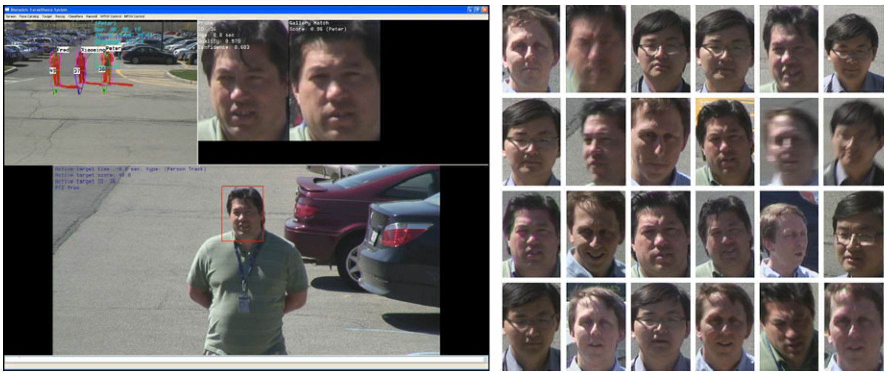


Fig. 14.1 On the left, a face recognition at a distance application showing tracked and identified subjects in wide field-of-view video (*upper left*), high-resolution narrow field-of-view video from an automatically controlled PTZ camera (*bottom*), and a detected and recognized facial image (*upper right*). On the right, some of the facial images captured by the system over a few minutes, selected to show the variation in facial image quality

The nature of the activity of subjects and the size of the coverage area can vary considerably with the application and this impacts the degree of difficulty. Subjects may be sparse and standing or walking along predictable trajectories, or they may be crowded, moving in a chaotic manner, and occluding each other. The coverage area may range from a few square meters at a doorway or choke point, to a transportation terminal, building perimeter, city block, or beyond. Practical solutions do involve image capture from a distance, but the field might be more accurately called *face recognition of noncooperative subjects over a wide area*. Figure 14.1 shows a FRAD system operating in a parking lot. There are two primary difficulties faced by FRAD. First, acquiring facial images from a distance. Second, recognizing the person in spite of imperfections in the captured data.

There are a wide variety of commercial, security, defense and marketing applications of FRAD. Some of the most important potential applications include:

- Access control: Unlock doors when cleared persons approach.
- Watch-list recognition: Raise an alert when a person of interest, such as a known terrorist, local offender or disgruntled ex-employee is detected in the vicinity.
- White-list recognition: Raise an alert whenever a person not cleared for the area is detected.
- Rerecognition: Recognize people recently imaged by a nearby camera for automatic surveillance with long-range persistent tracking.
- Event logging: For each person entering a region, catalog the best facial image.
- Marketing: Understand long-term store customer activities and behavior.

The *Handbook of Remote Biometrics* [53] also contains chapters on FRAD. The focus in that book is somewhat complementary, covering system issues and a more detailed look at illumination levels, optics and image sensors for face imaging at distances up to the 100–300 m range and beyond with both theoretical analysis and practical design advice.

Chapter 15
Face Recognition Using Near Infrared Images

Stan Z. Li and Dong Yi

15.1 Introduction

Face recognition should be based on intrinsic factors of the face, such as, the 3D shape and the albedo of the facial surface. Extrinsic factors that include illumination, eyeglasses, and hairstyle, are irrelevant to biometric identity, and hence their influence should be minimized. Out of all these factors, variation in illumination is a major challenge and needs to be tackled first.

Conventional visual (VIS) image based face recognition systems, academic and commercial, are compromised in accuracy by changes in environmental illumination, even for cooperative user applications in an indoor environment. In an in-depth study on influence of illumination changes on face recognition [1], Adini et al. examined several distance measures and local image operators, including Gabor filters, local directive filters, and edge maps, which were considered to be relatively insensitive to illumination changes for face recognition. Several conclusions were made: (i) lighting conditions, and especially light angle, drastically change the appearance of a face; (ii) when comparing unprocessed images, the changes between images of a person under different illumination conditions are larger than those between images of two persons under the same illumination; (iii) all the local filters under study, are not capable overcoming variations due to changes in illumination direction. The influence of illumination is also shown in evaluations such as Face Recognition Vendor Test [17].

Near infrared (NIR) based face recognition [11–14], as opposed to the conventional visible light (VIS) based methods, is an effective approach for overcoming the impact of illumination changes on face recognition. It uses a special purpose

S.Z. Li (✉) · D. Yi
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: szli@cbsr.ia.ac.cn

D. Yi
e-mail: dyi@cbsr.ia.ac.cn

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_15, © Springer-Verlag London Limited 2011

imaging device to capture front-lighted NIR face images [3–5], normalizing the illumination direction. Using a proper face feature representation, such as Local Binary Pattern (LBP) [2, 9, 18], variation in the illumination strength is also overcome. These lead to a complete illumination-invariant face representation. Problems caused by uncontrolled environmental illumination are minimized thereby, and difficulties in building the face matching engine are alleviated. The NIR approach usually achieves significantly higher performance than the VIS approach for cooperative user application scenarios in uncontrolled illumination environment. NIR face recognition products and integrated systems have been in the market and are used in many applications (refer to Chap. 1).

In this chapter, we introduce the NIR face recognition approach, describe the design of active NIR face imaging system, illustrate how to derive from NIR face image an illumination invariant face representation, and provide a learning based method for face feature selection and classification. Experiments are presented.

15.2 Active NIR Imaging System

The key aspect in the NIR face recognition approach is a special purpose NIR image capture hardware system [14]. Its goal is to overcome the problem arising from uncontrolled environmental light and produce face images of a good illumination condition for face recognition. Good illumination means (i) that the lighting to the face is from the frontal direction and (ii) that the face image has suitable pixel intensities.

To achieve illumination from the frontal direction, active NIR illuminators, for example, space light-emitting diodes (LEDs) around the camera lens, are used to illuminate the face from the front such that front-lighted NIR face images are acquired. This is similar to a camera with a flash light but the NIR lights work in the invisible spectrum of NIR, being nonintrusive to human eyes.

The following are the main requirements for the NIR imaging system:

1. The active NIR lights should be nonintrusive to human eyes.
2. The direction of the NIR lighting to the face should be fixed.
3. The active NIR light signals arriving at the camera sensor should override the signals from other light sources in the environment.

Here, the NIR lights mean the active NIR lights from the NIR imaging system, excluding NIR components in the environment such as sunlight and light bulbs.

This selective capture (of NIR light from the imaging system) can be achieved by the following methods:

1. Choose illuminators such as LEDs in an invisible spectrum. While a 850 nm LED light looks a dim dark red, 940 nm is entirely invisible.
2. Mount the NIR LEDs around the camera lens so as to illuminate from the frontal direction.

Chapter 16
Multispectral Face Imaging and Analysis

Andreas Koschan, Yi Yao, Hong Chang, and Mongi Abidi

16.1 Introduction

This chapter addresses the advantages of using multispectral narrow-band images for face recognition, as opposed to conventional broad-band images obtained by color or monochrome cameras (see also the chapter for a discussion of color in face analysis). Narrow-band images are by definition taken over a very small range of wavelengths, while broad-band images average the information obtained over a wide range of wavelengths. There are two primary reasons for employing multispectral imaging for face recognition.

First, we believe that there is distinctive facial information contained in certain narrow spectral bands which can be acknowledged and employed to enhance face recognition performance in comparison to broad-band color or black and white images. Broad-band imaging has the potential to degrade this information that is embedded in the narrow-band image due to the integration process over a wide range of wavelengths during the formation of the image.

Second, multispectral images can separate the illumination information from the reflectance of objects, so that we can use this illumination information to normalize the images. In contrast, it is nearly impossible to separate and employ the illumination distribution information from broad-band images. To verify the effectiveness of

A. Koschan (🖂) · H. Chang · M. Abidi
Imaging, Robotics, and Intelligent Systems Lab, University of Tennessee, Knoxville, TN 37996, USA
e-mail: akoschan@utk.edu

H. Chang
e-mail: hchang2@utk.edu

M. Abidi
e-mail: abidi@utk.edu

Y. Yao
Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY 12309, USA
e-mail: yi.yao@ge.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_16, © Springer-Verlag London Limited 2011

multispectral images for improving face recognition, two sequential procedures are taken into account: first, multispectral face image acquisition and second, spectral band selection.

To reduce information redundancy among multispectral images, complexity-guided distance-based band selection is introduced which uses a model selection criterion for an automatic selection. This selection can simplify the imaging process by reducing the number of multispectral images to be taken under a given illumination. In other words, the goal is to identify a small set of optimal multispectral bands to be taken under a given illumination as opposed to acquiring a large set of multispectral bands over the entire visible spectrum.

The performance of selected bands outperforms the conventional images by up to 15%. From the significant performance improvement via complexity-guided distance-based band selection, we conclude that specific facial information carried in certain narrow-band spectral images can enhance face recognition performance compared to broad-band images. In addition, the algorithm is equally useful and successful in a wide variety of recognition schemes.

16.2 Multispectral Imaging

Multispectral imaging is a technique that provides images of a scene at multiple wavelengths and can generate precise optical spectra at every pixel. A *multispectral image* is a collection of several monochrome images of the same scene, each of them taken with additional receptors sensitive to other frequencies of the visible light, or to frequencies beyond visible light, like the infrared region of electromagnetic continuum. Each image is referred to as a *band* or a *channel*. Multispectral imaging produces a three-dimensional image cube with two spatial dimensions (horizontal and vertical) and one spectral dimension. The spectral dimension contains spectral information for each pixel on the multispectral cube. A *multispectral image* can be represented as

$$C(x, y) = \big(\mu_1(x, y), \mu_2(x, y), \dots, \mu_{N_B}(x, y)\big)^{\mathrm{T}} (\mu_1, \mu_2, \dots, \mu_{N_B})^{\mathrm{T}}. \tag{16.1}$$

The signal strength $u_k(x, y)$ of a camera sensor in a certain wavelength range, $\lambda_{\min}$ to $\lambda_{\max}$, can be represented as

$$u_k(x, y) = \int_{\lambda_{\min}}^{\lambda_{\max}} R(x, y, \lambda) L(x, y, \lambda) S_k(x, y, \lambda)\, d\lambda, \tag{16.2}$$

with $k = 1, \dots, N_B$, where $N_B = 1$ for monochromatic images and $N_B = 3$ for three-channel color images. The parameters (x, y) indicate the pixel location in the image. $R(x, y, \lambda)$ is the spectral surface reflectance of the object, $L(x, y, \lambda)$ is the spectral distribution of the illumination, and $S_k(x, y, \lambda)$ is the spectral sensitivity of the camera corresponding to channel k. The entire possible integration wavelength range can be in the visible spectrum, 400–720 nm, or in addition may include infrared spectrum depending on the camera design.

Chapter 17
Face Recognition Using 3D Images

I.A. Kakadiaris, G. Passalis, G. Toderici, E. Efraty, P. Perakis, D. Chu, S. Shah, and T. Theoharis

17.1 Introduction

Our face is our password—face recognition promises to revolutionize the way we identify individuals in a nonintrusive and convenient manner. Even though research in face recognition has spanned over nearly three decades, only 2D systems, with limited adoption to practical applications, have been developed so far. The primary reason behind this is the low accuracy of 2D face recognition systems in the presence of: (i) pose variations between the gallery and probe datasets, (ii) variations in lighting, and (iii) variations in the presence of expressions and/or accessories. The above conditions generally arise when noncooperative subjects are involved, which is the very case that demands accurate recognition.

Face recognition using 3D images was introduced in order to overcome these challenges. It was partly made possible by significant advances in 3D scanner technology. However, even 3D face recognition has faced significant challenges which have hindered its adoption for practical applications. The main problem of 3D face recognition is the high cost and fragility of 3D scanners. Over the last seven years, our research team has focused on exploring the usefulness of 3D data and the development of models for face recognition (under the general name URxD).

In this chapter, we present advances that aid in overcoming the challenges encountered in 3D face recognition. First, we present a fully automatic 3D face recognition system, UR3D, which has been proven to be robust under variations in expressions. The fundamental idea of this system is the description of facial data using an Annotated Face Model (AFM). The AFM is fitted to the facial scan using

I.A. Kakadiaris (✉) · G. Passalis · G. Toderici · E. Efraty · P. Perakis · D. Chu · S. Shah · T. Theoharis
Computational Biomedicine Lab, Department of Computer Science, University of Houston, Houston, TX 77204, USA
e-mail: ioannisk@grip.cis.upenn.edu

G. Passalis · P. Perakis · T. Theoharis
Computer Graphics Laboratory, Department of Informatics and Telecommunications, University of Athens, Ilisia 15784, Greece

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_17, © Springer-Verlag London Limited 2011

a subdivision-based deformable model framework. The deformed model captures the details of an individual's face and represents this 3D geometry information in an efficient 2D representation by utilizing the model's parametrization. This representation is analyzed in the wavelet domain and the associated wavelet coefficients define the metadata that are used for comparing the different subjects. These metadata are both compact and descriptive. This approach that involves geometric modeling of the human face allows greater flexibility, better understanding of the face recognition issues, and requires no training.

Second, we demonstrate how pose variations are handled in 3D face recognition. The 3D scanners that are used to obtain facial data are usually nonimmersive which means that only a partial 3D scan of the human face is obtained, particularly so in noncooperative, practical conditions. Thus, there are often missing data of the frontal part of the face. This can be overcome by identifying a number of landmarks on each 3D facial scan thereby allowing correct registration with the AFM, independent of the original pose of the face. For nonfrontal scans, missing data can be added by exploiting facial symmetry, assuming that at least half of the face is visible. This is achieved by improving the subdivision-based deformable model framework to allow symmetric fitting. Symmetric fitting alleviates the missing data problem and facilitates the creation of geometry images that are pose invariant. Another alternative to tackle the missing data problem is to attempt recognition based on the facial profile; this approach is particularly useful in recognizing car drivers from side view images. In this approach, the gallery includes facial profile information under different poses, collected from subjects during enrollment. These profiles are generated by projecting the subjects' 3D face data. Probe profiles are extracted from the input images and compared to the gallery profiles.

Finally, we demonstrate how the problems related to the cost of 3D scanners can be mitigated through hybrid systems. Such systems employ 3D scanners for the enrollment of subjects, which can take place in a few specialized locations, and 2D cameras at points of authentication, which can be multiple and dispersed. It is practical to adopt this approach if hybrid systems can improve the accuracy of a 2D system. During enrollment, 2D+3D data (2D texture and 3D shape) are used to build subject-specific annotated 3D models. To achieve this, an AFM is fitted to the raw 2D+3D data using a subdivision-based deformable framework. A geometry image representation is then extracted using the parametrization of the model. During the verification phase, a single 2D image is used as the input to map the subject-specific 3D AFM. Given the pose in the 2D image, an Analytical Skin Reflectance Model (ASRM) is then applied to the gallery AFM to transfer the lighting from the probe to the texture in the gallery. The matching score is computed using the relit gallery texture and the probe texture. This hybrid method surpasses the accuracy of 2D face recognition system in difficult datasets.

Chapter 18
Facial Action Tracking

Jörgen Ahlberg and Igor S. Pandzic

18.1 Introduction

The problem of facial action tracking has been a subject of intensive research in the last decade. Mostly, this has been with such applications in mind as face animation, facial expression analysis, and human-computer interfaces. In order to create a face animation from video, that is, to capture the facial action (facial motion, facial expression) in a video stream, and then animate a face model (depicting the same or another face), a number of steps have to be performed. Naturally, the face has to be detected, and then some kind of model has to be fitted to the face. This can be done by aligning and deforming a 2D or 3D model to the image, or by localizing a number of facial landmarks. Commonly, these two are combined. The result must in either case be expressed as a set of parameters that the face model in the receiving end (the face model to be animated) can interpret.

Depending on the application, the model and its parameterization can be simple (e.g., just an oval shape) or complex (e.g., thousands of polygons in layers simulating bone and layers of skin and muscles). We usually wish to control appearance, structure, and motion of the model with a small number of parameters, chosen so as to best represent the variability likely to occur in the application. We discriminate here between rigid face/head tracking and tracking of facial action. The former is typically employed to robustly track the faces under large pose variations, using a rigid face/head model (that can be quite non-face specific, e.g., a cylinder). The latter here refers to tracking of facial action and facial features, such as lip and eyebrow

J. Ahlberg (✉)
Division of Information Systems, Swedish Defence Research Agency (FOI), P.O. Box 1165,
583 34 Linköping, Sweden
e-mail: jorahl@foi.se

I.S. Pandzic
Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb,
Croatia
e-mail: igor.pandzic@fer.hr

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_18, © Springer-Verlag London Limited 2011

motion. The treatment in this chapter is limited to tracking of facial action (which, by necessity, includes the head tracking).

The parameterization can be dependent or independent on the model. Examples of model independent parameterizations are MPEG-4 Face Animation Parameters (see Sect. 18.2.3) and FACS Action Units (see Sect. 18.2.2). These parameterizations can be implemented on virtually any face model, but leaves freedom to the model and its implementation regarding the final result of the animation. If the transmitting side wants to have full control of the result, more dense parameterizations (controlling the motion of every vertex in the receiving face model) are used. Such parameterizations can be implemented by MPEG-4 Facial Animation Tables (FAT), morph target coefficients, or simply by transmitting all vertex coordinates at each time step.

Then, the face and its landmark/features/deformations must be tracked through an image (video) sequence. Tracking of faces has received significant attention for quite some time but is still not a completely solved problem.

18.1.1 Previous Work

A plethora of face trackers are available in the literatures, and only a few of them can be mentioned here. They differ in how they model the face, how they track changes from one frame to the next, if and how changes in illumination and structure are handled, if they are susceptible to drift, and if real-time performance is possible. The presentation here is limited to monocular systems (in contrast to stereo-vision) and 3D tracking.

18.1.1.1 Rigid Face/Head Tracking

Malciu and Prêteux [36] used an ellipsoidal (alternatively an ad hoc Fourier synthesized) textured wireframe model and minimized the registration error and/or used the optical flow to estimate the 3D pose. LaCascia et al. [31] used a cylindrical face model with a parameterized texture being a linear combination of texture warping templates and orthogonal illumination templates. The 3D head pose was derived by registering the texture map captured from the new frame with the model texture. Stable tracking was achieved via regularized, weighted least-squares minimization of the registration error.

Basu et al. [7] used the Structure from Motion algorithm by Azerbayejani and Pentland [6] for 3D head tracking, refined and extended by Jebara and Pentland [26] and Ström [56] (see below). Later rigid head trackers include notably the works by Xiao et al. [64] and Morency et al. [40].

Chapter 19
Facial Expression Recognition

Yingli Tian, Takeo Kanade, and Jeffrey F. Cohn

19.1 Introduction

Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. Facial expression analysis has been an active research topic for behavioral scientists since the work of Darwin in 1872 [21, 26, 29, 83]. Suwa et al. [90] presented an early attempt to automatically analyze facial expressions by tracking the motion of 20 identified spots on an image sequence in 1978. After that, much progress has been made to build computer systems to help us understand and use this natural form of human communication [5, 7, 8, 17, 23, 32, 43, 45, 57, 64, 77, 92, 95, 106–108, 110].

In this chapter, facial expression analysis refers to computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from visual information. Sometimes the facial expression analysis has been confused with emotion analysis in the computer vision domain. For emotion analysis, higher level knowledge is required. For example, although facial expressions can convey emotion, they can also express intention, cognitive processes, physical effort, or other intra- or interpersonal meanings. Interpretation is aided by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing [11, 79, 80]. Computer facial expression analysis systems need to analyze the facial actions regardless of context, culture, gender, and so on.

Y. Tian (✉)
Department of Electrical Engineering, The City College of New York, New York, NY 10031, USA
e-mail: ytian@ccny.cuny.edu

T. Kanade
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: tk@cs.cmu.edu

J.F. Cohn
Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: jeffcohn@pitt.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_19, © Springer-Verlag London Limited 2011

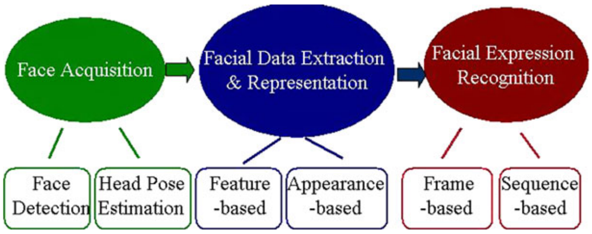


Fig. 19.1 Basic structure of facial expression analysis systems

The accomplishments in the related areas such as psychological studies, human movement analysis, face detection, face tracking, and recognition make the automatic facial expression analysis possible. Automatic facial expression analysis can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human computer interface (HCI).

19.2 Principles of Facial Expression Analysis

19.2.1 Basic Structure of Facial Expression Analysis Systems

Facial expression analysis includes both measurement of facial motion and recognition of expression. The general approach to automatic facial expression analysis (AFEA) consists of three steps (Fig. 19.1): face acquisition, facial data extraction and representation, and facial expression recognition.

Face acquisition is a processing stage to automatically find the face region for the input images or sequences. It can be a detector to detect face for each frame or just detect face in the first frame and then track the face in the remainder of the video sequence. To handle large head motion, the head finder, head tracking, and pose estimation can be applied to a facial expression analysis system.

After the face is located, the next step is to extract and represent the facial changes caused by facial expressions. In facial feature extraction for expression analysis, there are mainly two types of approaches: geometric feature-based methods and appearance-based methods. The geometric facial features present the shape and locations of facial components (including mouth, eyes, brows, nose, etc.). The facial components or facial feature points are extracted to form a feature vector that represents the face geometry. With appearance-based methods, image filters, such as Gabor wavelets, are applied to either the whole-face or specific regions in a face image to extract a feature vector. Depending on the different facial feature extraction methods, the effects of in-plane head rotation and different scales of the faces can be eliminated by face normalization before the feature extraction or by feature representation before the step of expression recognition.

Facial expression recognition is the last stage of AFEA systems. The facial changes can be identified as facial action units or prototypic emotional expressions (see Sect. 19.3.1 for definitions). Depending on whether the temporal information is

Chapter 20
Face Synthesis

Yang Wang, Zicheng Liu, and Baining Guo

20.1 Introduction

How to synthesize photorealistic images of human faces has been a fascinating yet difficult problem in computer graphics. Here, the term "face synthesis" refers to synthesis of still images as well as synthesis of facial animations. In general, it is difficult to draw a clean line between the synthesis of still images and that of facial animations. For example, the technique of synthesizing facial expression images can be directly used for generating facial animations, and most of the facial animation systems involve the synthesis of still images. In this chapter, we focus more on the synthesis of still images and skip most of the aspects that mainly involve the motion over time.

Face synthesis has many interesting applications. In the film industry, people would like to create virtual human characters that are indistinguishable from the real ones. In games, people have been trying to create human characters that are interactive and realistic. There are commercially available products [18, 19] that allow people to create realistic looking avatars that can be used in chat rooms, e-mail, greeting cards, and teleconferencing. Many human-machine dialog systems use realistic-looking human faces as visual representation of the computer agent that interacts with the human user. Face synthesis techniques have also been used for talking head compression in the video conferencing scenario.

Y. Wang (✉)
Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: wangy@cs.cmu.edu

Z. Liu
Microsoft Research, Redmond, WA 98052, USA
e-mail: zliu@microsoft.com

B. Guo
Microsoft Research Asia, Beijing 100080, China
e-mail: bainguo@microsoft.com

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_20, © Springer-Verlag London Limited 2011

The techniques of face synthesis can be useful for face recognition too. Romdhani et al. [47, 48] used their three dimensional (3D) face modeling technique for face recognition with different poses and lighting conditions. Qing et al. [44] used the face relighting technique as proposed by Wen et al. [59] for face recognition under a different lighting environment. Wang et al. [57] used the 3D spherical harmonic morphable model (SHBMM), an integration of spherical harmonics into the morphable model framework, for face recognition under arbitrary pose and illumination conditions. Many face analysis systems use an analysis-by-synthesis loop where face synthesis techniques are part of the analysis framework.

In this chapter, we review recent advances on face synthesis including 3D face modeling, face relighting, and facial expression synthesis.

20.2 Face Modeling

In the past a few years, there has been a lot of work on the reconstruction of face models from images [12, 23, 27, 41, 47, 52, 67]. There are commercially available software packages [18, 19] that allow a user to construct their personalized 3D face models. In addition to its applications in games and entertainment, face modeling techniques can also be used to help with face recognition tasks especially in handling different head poses (see Romdhani et al. [48] and Chap. 10). Face modeling techniques can be divided into three categories: face modeling from an image sequence, face modeling from two orthogonal views, and face modeling from a single image. An image sequence is typically a video of someone's head turning from one side to the other. It contains a minimum of two views. The motion between each two consecutive views is relatively small, so it is feasible to perform image matching.

20.2.1 Face Modeling from an Image Sequence

Given an image sequence, one common approach for face modeling typically consists of three steps: image matching, structure from motion, and model fitting. First, two or three relatively frontal views are selected, and some image matching algorithms are used to compute point correspondences. The selection of frontal views are usually done manually. Point correspondences are computed either by using dense matching techniques such as optimal flow or feature-based corner matching. Second, one needs to compute the head motion and the 3D structures of the tracked points. Finally, a face model is fitted to the reconstructed 3D points. People have used different types of face model representations including parametric surfaces [13], linear class face scans [5], and linear class deformation vectors [34].

Fua and Miccio [13, 14] computed dense matching using image correlations. They then used a model-driven bundle adjustment technique to estimate the motions and compute the 3D structures. The idea of the model-driven bundle adjustment is to add a regularizer constraint to the traditional bundle adjustment formulation. The

Part III
Performance Evaluation: Machines and Humans

Chapter 21
Evaluation Methods in Face Recognition

P. Jonathon Phillips, Patrick Grother, and Ross Micheals

21.1 Introduction

Face recognition from still frontal images has made great strides over the last twenty years. Over this period, error rates have decreased by three orders of magnitude when recognizing frontal face in still images taken with consistent controlled illumination in an environment similar to a studio [5, 10, 15, 17–22]. Under these conditions, error rates below 1% at a false accept rate of 1 in 1000 were reported in the Face Recognition Vendor Test[1] (FRVT) 2006 and the Multiple Biometric Evaluation (MBE) 2010 [10, 21].

The heart of designing and conducting evaluations is the experimental protocol. The protocol states how an evaluation is to be conducted and how the results are to be computed. In this chapter, we concentrate on describing the FERET and FRVT 2002 protocols. The FRVT 2002 evaluation protocol is based on the FERET evaluation protocols. The FRVT 2002 protocol is designed for biometric evaluations in general, not just for evaluating face recognition algorithms. These two evaluation protocols served as a basis for the FRVT 2006 and MBE 2010 evaluations.

The FRVT 2002 protocol was designed to allow for computing a wide range of performance statistics. This includes the standard performance tasks of open-set and closed-set identification, and verification. It also allows for resampling techniques,

[1] Performance results in this chapter are labeled by the test participants. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

P.J. Phillips (✉) · P. Grother · R. Micheals
National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
e-mail: jonathon@nist.gov

P. Grother
e-mail: pgrother@nist.gov

R. Micheals
e-mail: rossm@nist.gov

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_21, © Springer-Verlag London Limited 2011

similarity score normalization, measuring the variability of performance statistics, and covariate analysis [1–4, 9, 11, 14].

21.2 Performance Measures

In face recognition and biometrics, performance is reported on three standard tasks: verification, open-set and closed-set identification. Each task has its own set of performance measures. All three tasks are closely related, with open-set identification being the general case.

A biometric system works by processing biometric samples. *Biometric samples* are recordings of a feature of a person that allows that person to be recognized. Examples of biometric samples are facial images and fingerprints. A biometric sample can consist of multiple recordings, for example, five images of a person acquired at the same time or a facial image and a fingerprint.

Computing performance requires three sets of images. The first is a *gallery* $\mathcal{G}$, which contains biometric samples of the people known to a system. The other two are *probe sets*. A *probe* is a biometric sample that is presented to the system for recognition, where recognition can be verification or identification. The first probe set is $\mathcal{P}_{\mathcal{G}}$ that contains biometric samples of people in a gallery (these samples are different from the samples in the gallery). The other probe set is $\mathcal{P}_{\mathcal{N}}$, which contains biometric samples of people that are not in a gallery.

Closed-set identification is the classic performance measure used in the automatic face recognition community, where it is known as identification. In closed-set identification, the basic question asked is: whose face is this? This question is meaningful for closed-set identification, since the biometric sample in a probe is always someone in the gallery. The general case of closed-set identification is open-set identification.

In open-set identification, the person in the probe does not have to be somebody in the gallery. In open-set identification, the basic question asked is: do we know this face? In open-set identification, a system has to decide if the probe contains an image of a person in the gallery. If a system decides that a person is in the gallery, then the system has to report the identity of the person. When the gallery is small, open-set identification can be referred to as a watch list task. When the gallery is large, then open-set identification models mugshot book searching and the operation of large automatic fingerprint identification systems (AFIS as they are sometimes called). Open-set and closed-set identification are sometimes referred to as 1 to many matching or 1:N matching. Depending on the context and author, 1 to many matching or 1:N matching can refer to either open-set or closed-set identification.

In a verification task, a person presents a biometrics sample to a system and claims an identity. The system has to decide if the biometric sample belongs to the claimed identity. In verification, the basic question asked is: is this person who he claims to be? Verification is also called authentication or 1 to 1 matching.

Chapter 22
Dynamic Aspects of Face Processing in Humans

Heinrich H. Bülthoff, Douglas W. Cunningham, and Christian Wallraven

22.1 Introduction

The human face is capable of a wide variety of facial expressions that manifest themselves as usually highly non-rigid deformations of the face. On the one hand, this presents the visual system with a problem: Recognizing someone requires determining what information in the face remains constant despite the various facial deformations. Extraction of such invariant features will allow me, for example, to identify my neighbor regardless of whether he or she is smiling or looking sad. On the other hand, the impressive repertoire of changes can also be seen as a positive: It provides considerable information. The particular way my neighbor smiles or looks sad might well be used for identification, similar to how Jack Nicholson's and Tom Cruise's smiles are very specific to them.

In addition to potentially providing information about who someone is, facial deformations can help us to infer something about a person's age, social status, general health, level of fatigue, and focus of attention. Likewise, changes in the facial surface play a central, albeit often ignored, role in communication. Facial deformations that serve this latter role are generally referred to as facial expressions.

A distinction should be drawn between the information that *is* present in a *specific image* and the information that *must be* present for that expression or person to be

H.H. Bülthoff (✉) · D.W. Cunningham · C. Wallraven
Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany
e-mail: heinrich.buelthoff@tuebingen.mpg.de

H.H. Bülthoff · C. Wallraven
Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

C. Wallraven
e-mail: wallraven@korea.ac.kr

D.W. Cunningham
Brandenburg Technical University, 03046 Cottbus, Germany
e-mail: douglas.cunningham@tu-cottbus.de

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_22, © Springer-Verlag London Limited 2011

recognized. Trying to determine what information is perceptually necessary not only provides critical insights into how humans process faces, but can also yield clues for the design of automated facial recognition and synthesis systems.

Almost all research on the perception of faces—both for identity and expression—has tended to focus on the relatively stable aspects of faces. Some of this information is invariant to deformations of the facial surface, such as the color of or distance between the eyes. In other cases, the result of the deformation *is* the information, such as the shape of the mouth. In such cases, usually the maximum deformation (or peak expression) is examined. In other words, there is a pervasive emphasis on *static* facial information. To some degree, this focus on static information is driven by technology. It is difficult to systematically manipulate a photograph in order to provide the systematic and parameterized variations needed for perceptual experiments without making the photograph look unrealistic. It is considerably more difficult to perform the same manipulations on a *sequence* of images without introducing artifacts either in any given image or across images [29, 80].

In general, however, human faces are not static entities. Indeed, if we meet someone who never moved their face, we would most likely be rather uncomfortable. Some have gone so far as to argue that an individual (specifically, an android) that looks like a human but moves either incorrectly or not at all (i.e., has a "dead" face) will—as a result of the zombie-like appearance—lead to humans being repulsed by that individual [75]. This hypothesis is referred to as the "uncanny valley". Regardless of whether a zombie—or zombie-like individual—will repulse humans or not, it is clear that the pattern of change over time is itself a great source of information for many visual processes [46] and can often be used to discern real from synthesized stimuli [106].

Fortunately, recent advances in technology, have allowed researchers to carefully and systematically alter video sequences without introducing noticeable artifacts and thus begin to examine the role of motion in face processing. Before one can determine what types of motion are used (i.e., uncover the dynamic features), one must determine if motion plays any role at all. It has been shown that facial motion can provide information about gender [15, 51] and age [14].

In this chapter, we will focus on the role of motion in identity (Sect. 22.2) and expression (Sect. 22.3) recognition in humans, and explain its developmental and neurophysiological aspects. We will make some inferences and conclusions based on results from literature.

22.2 Dynamic Information for Identity

Correctly identifying other people is critical to everyday survival and there has been a considerable amount of research on how such a task might be performed. The literature on how humans use faces to recognize people is quite extensive (for a review, see Chap. 26, this volume or [91]). While the great majority of this literature has focused on static information for identity, there has been an increasing interest in dynamic information (see, e.g., [81]).

Chapter 23
Face Recognition by Humans and Machines

Alice J. O'Toole

23.1 Introduction

We recognize individual human faces dozens of times each day, with seemingly minimal effort. Our repertoire of known faces includes those of our family, friends, coworkers, acquaintances, and the many familiar faces we see in the news and entertainment media. At its best, human face recognition is highly robust to changes in viewing angle and illumination. It is also robust across the set of nonrigid face deformations that define emotional expressions and speech movements. Humans can even manage, in some cases, to recognize faces over decade-long lapses between encounters. Over time spans early in life, the head structure and facial features change markedly as children grow into adolescents and adults. Later in life, faces can age in ways that alter both the shape of the facial surface and the texture of the skin, as well as the color of the hair and eyebrows. By almost any current measure of the complexity of the computational vision problems solved to accomplish this task, the human face recognition system is impressive.

The description of human face recognition skills just offered is a *best case scenario* and one that is true only for the faces of people we know reasonably well (e.g., friends and family) or have seen many times in the popular media (e.g., Barack Obama, Angelina Jolie). For the faces of people we know from only a single or small number of encounters, human performance is not similarly robust. This is abundantly clear from the mundane day-to-day mistakes we make, and more critically, from the many well-documented cases of the fallibility of eyewitness identifications. In this latter case, a person may be seen initially under sub-optimal viewing conditions. A witness may then be asked days, weeks, or even months later to make an identification. Human recognition can be highly prone to error in these cases [9, 10].

A.J. O'Toole (✉)
School of Behavioral and Brain Sciences, The University of Texas at Dallas,
800 W. Campbell Rd., Richardson, TX 75083-0688, USA
e-mail: otoole@utdallas.edu

S.Z. Li, A.K. Jain (eds.), *Handbook of Face Recognition*,
DOI 10.1007/978-0-85729-932-1_23, © Springer-Verlag London Limited 2011

A key issue both for computer vision researchers and for psychologists is to understand how it is possible to create a representation of faces that achieves the kind of robust face recognition people show when they know someone well. Understanding the changes that occur in the quality of a face representation as a newly learned face becomes increasingly familiar over time may offer insight into the critical problems that still challenge even the best available computer-based face recognition systems. It may also help us to understand the conditions under which human face recognition is reliable. We will conclude, ultimately, that the best current face recognition algorithms are well on their way to achieving the recognition ability humans show for unfamiliar faces, but have a long way to go before they can compete with humans in the best case scenario.

In the first part of this chapter, I will describe the characteristics of human face processing, emphasizing psychological studies that illustrate the nature of human face representations. We begin with a brief overview of the multiple tasks humans do with faces, including identification, categorization, and expression perception. Next we will look at some characteristics of the human representation of faces that distinguish it from representations used in machine vision. In particular, we focus on the advantages of norm-based coding for identification tasks. This optimizes the feature dimensions used to code faces, but may have a cost in generalization to faces that are not described by the derived sets of features (e.g., faces of "other" races or ethnicities).

In the second part of the chapter, I will discuss a series of recent studies that compare human performance to the performance of state-of-the-art face recognition systems. One goal of these comparisons is to establish human benchmarks for the performance of face recognition algorithms. A second goal is to understand the strengths and weaknesses of humans and machines at the task of face recognition and to come up with strategies for optimally combining human and machine recognition decisions. A third aim of these studies is to understand qualitative similarities and differences in the pattern of errors made by humans and machines. We will argue that studying human face recognition in this way can help us to anticipate and mitigate the errors made by both human and computer-based systems.

23.2 What Humans Do with Faces

Perhaps the most remarkable aspect of the human face is the diversity of information it provides simultaneously to the human observer for solving different "tasks". These tasks include recognition, identification, visually-based categorization (e.g., sex, race, age), and emotional/facial expression perception. The challenge for the human system is to extract and apply the information needed for the task at hand. As we will see, the coexistence of identity information (e.g., distinctive facial features) and social information (e.g., smiles, head turns, etc.) in the human face makes the problem of recognition even more challenging.