Chapter 4

Modeling, Reconstruction and Tracking for Face Recognition

With the development of biometric techniques, automatic identity control systems have been invested in many places and facilities (e.g. airports and secure premises) during the last years. To improve the traffic flow at these recognition systems, it is necessary to minimize as much as possible the constraints imposed on the user. To meet this objective, it is necessary to perform "on-the-fly" acquisitions, without constraining the user to stop and stand in front of a sensor. In this chapter, we therefore focus on the use of facial biometrics, and more specifically on issues related to on-the-fly face acquisition. To enable the authentication within such systems, we have to solve a number of issues related to the facial shape and texture estimation. To address the theoretical aspects related to face acquisition and reconstruction, we consider the following framework: a multiview acquisition system is placed at the entrance of a room, a corridor, or a car park, for instance, and the aim is to identify or authenticate the person observed through this device.

4.1. Background

The requirements of a biometric system are varied, for example its ease of use, its speed of execution, its non-intrusiveness to users, its cost, and its reliability. The first three points are essential for systems designed for the general public, where the number of users is important and where these users are not specifically trained to

Chapter written by Catherine HEROLD, Vincent DESPIEGEL, Stéphane GENTRIC, Séverine DUBUISSON and Isabelle BLOCH.

use these systems (and should not have to be). For instance, in the case of passengers at airports, many people will have to use a biometric system a limited number of times during their lives. An effective way to increase the usability and fluidity of this biometric system is to minimize as much as possible the constraints on the user's behavior during the acquisition process. As no specific action is required from the user, there is no possible mistake from him/her during the acquisition, thus reducing the required time.

The various available sources of biometric information (e.g. fingerprints, irises, faces and veins) do not have the same requirements. The common fingerprints or iris biometrics require a static position during the acquisition. They are also less accepted by users than facial biometrics, which is more natural to humans. For face acquisition, it is easy to imagine a protocol without contact or immobilization constraint, making it a biometrics both faster and much more accepted. The user does not experience the need to cooperate during the acquisition.

4.1.1. *Applications of face recognition*

In recent years, the growth of facial biometrics has been particularly important. It is, indeed, used for many purposes:

- for entrance or secure access control (identification relative to a database of authorized persons);
- for border control (authentication with passport);
- for right delivery (voter card, driving license, benefits, etc.);
- for police investigation.

In all these applications, facial biometrics can be used alone or in conjunction with other biometrics.

4.1.2. *On-the-fly authentication*

Many face acquisition systems require a specific behavior from the user, such as their immobilization in front of one or more cameras. This constraint significantly slows down the process of identity checking.

The main reason for this constraint is that the majority of facial biometric systems are based on comparisons between two views under the same pose to establish a matching score. The recorded reference views of individuals are usually frontal views (passport photo). The aim of the acquisition system is to provide a similar frontal view to proceed to the comparison. For systems where the user must stop in front of the sensor, it is fairly easy to acquire this type of view directly. However, if the acquisition is unconstrained, the face is seen under various poses. The frontal view should then be generated from observations to make the correspondence verification. This step is called "frontalization".

Other methods of comparison are also possible, as in [VET 97], where the author relaxes the conditions of pose similarity with generated views under new poses using computer graphics methods. Two views can also be compared through the three-dimensional (3D) shape and texture parameters that are estimated on each of them [BLA 03b]. Finally, there are also methods based on video streams that analyze facial dynamics to identify an individual, in addition to facial appearance [MAT 09]. However, in this chapter, we limit ourselves to a comparison between two frontal views, which corresponds to the majority of scenarios involving a passport photo.

To obtain the frontal view of an observed face, the general idea is to first estimate its 3D reconstruction (shape and texture) and then to generate the corresponding frontal view. The pose, shape, texture, and lighting estimation, which leads to the step of frontal view generation, is the core of this chapter. To evaluate these parameters, many acquisition systems are available. We limit the scope of this chapter to approaches that rely solely on video acquisitions made by common cameras. Other methods also exist, but they require more complete (3D scanners, depth sensors [ZOL 11]), or more intrusive (markers on the face [HUA 11] and structured light projection [ZHA 04]) systems and are therefore not discussed here.

Even with a multicamera system, a wide variety of information is available to reconstruct the 3D face: the system calibration and 3D models of faces, for example. We review these types of information in section 4.2 before detailing the approaches based on one or more views simultaneously acquired in sections 4.3 (geometric approaches), 4.4 (model-based approaches), and 4.5 (hybrid approaches). Finally, in section 4.6, we detail the approaches that integrate the time information using specifically video inputs.

To provide an overall view of the process, here is an example of an on-the-fly facial acquisition system (see Figure 4.1). No specific interaction from the user is required in order to accelerate the whole process of authentication (or identification). During his visit, the user's head is tracked in the general 3D coordinate system and the head model parameters are estimated from different available views (Figure 4.2(a)) in order to match as best as possible the face of the tracked person. At each moment, new observations are available and the face model of the individual can be computed or updated. New views, particularly the frontal view (Figure 4.2(b)), can then be generated in order to compare it with a database (identification) or to a passport photo (authentication). The whole process is summarized in Figure 4.2(c) and detailed in [MOË 10].

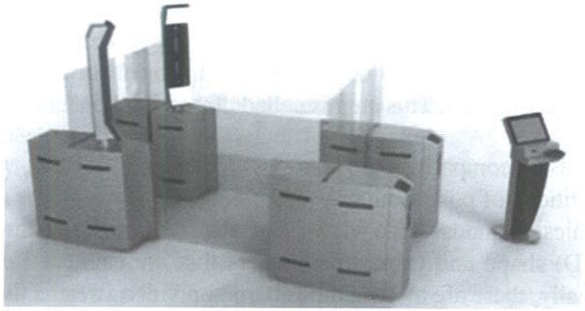
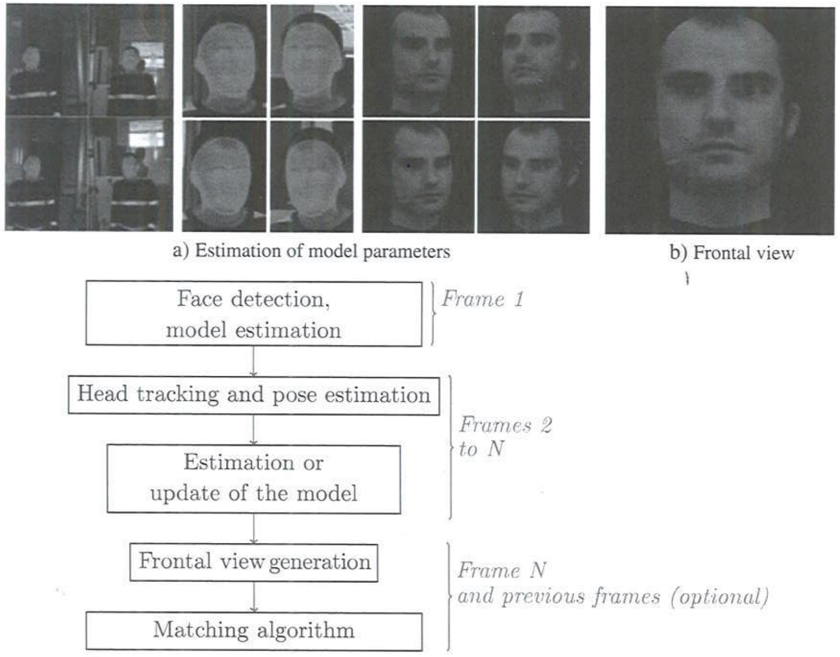


Figure 4.1. *On-the-fly facial authentication system*



a) Estimation of model parameters

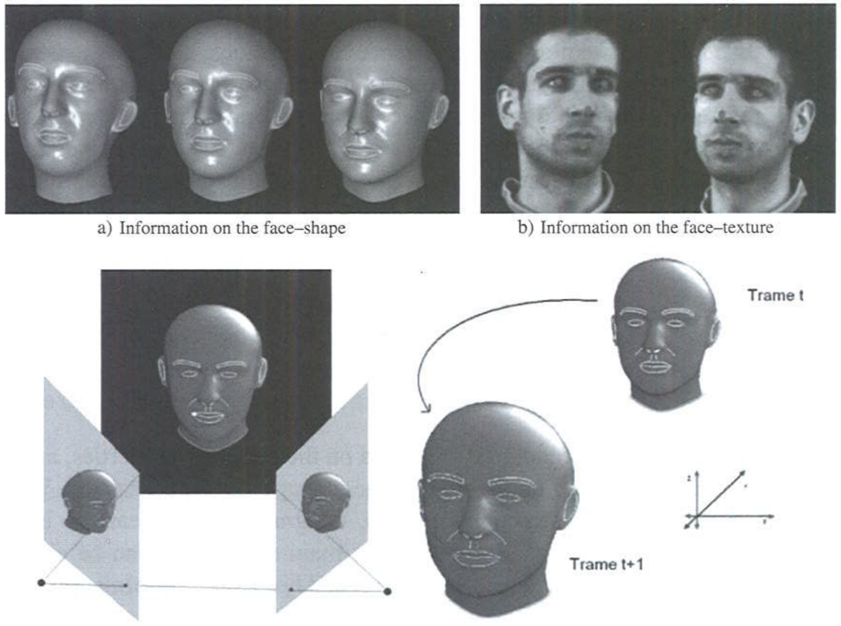
b) Frontal view

c) Overall processing workflow

Figure 4.2. *Overall process of tracking and authentication (Source: [HER 11])*

4.2. Types of available information

From a set of synchronized videos, much information can be used to reconstruct a frontal view of the observed face. Here, we distinguish between two types of data: the first is related to the properties of the acquisition system and the second is related to the nature of the object to be reconstructed, i.e. the face.



a) Information on the face–shape

b) Information on the face–texture

c) Information of the system–geometric constraints

d) Tracking–time constraints

Figure 4.3. *Available information to estimate the face from video acquisitions*

4.2.1. *Information related to the acquisition system*

By using a multicamera system, it is possible to rely on a set of synchronized views in order to match two-dimensional (2D) points and estimate the associated 3D points. Furthermore, if the system calibration is known, the epipolar constraints will allow one to improve the point matching between views. Many methods have been developed to estimate the calibration parameters of a single (or a set of) camera(s), with or without calibration pattern [HAR 04, ZHA 00]. Geometric constraints induced by the calibration thus allow one to reconstruct the shape of an object. Another solution, proposed in some algorithms, is to jointly estimate the calibration of the system and the position of the matched points.

Furthermore, if the system is installed in a controlled environment (with known light position and orientation), the shading of the object can be used to reconstruct it by a technique known as “shape from shading” [ZHA 99]. However, it is not always possible to control the light environment and to know the light source properties. By adding some assumptions on the shape properties of the object, this method can be applied to estimate its surface.

Finally, temporal information can also be used to reconstruct the observed face. It is first beneficial to exploit the coherence of positions and poses estimated between successive times in a tracking process. Moreover, the reconstruction can be made from different views of the same video stream by optimizing jointly the shape and the pose of the object, using correlation between successive views. This technique, sometimes called structure from motion, has already been used for many applications: urban environments reconstruction filmed from a vehicle, buildings reconstruction [POL 04], objects observed by a mobile webcam [NEW 10], etc. We describe at the end of this chapter how to use the video stream to consolidate facial reconstruction.

4.2.2. *Facial features*

All the aforementioned techniques are based on the system’s properties, and they consider no *prior* information on the type of the object to be reconstructed. We will now focus on the 3D reconstruction of faces, integrating facial features into the process. Depending on the type of features, the main approaches can be classified into two categories: texture information or color, and head shape information.

In the first category, we can differentiate global descriptors, which specify the properties of a face as a whole, and local descriptors, which locally describe some facial parts or feature points (such as the MPEG-4 FACE norm [PAN 03]). Haar wavelets and Gabor filters are two examples of descriptors commonly used to characterize a face or one of its parts. These descriptors are used in detection algorithms that identify the positions of faces or points of interest within an image. Descriptors are usually built from positive and negative training sets to find discriminative characteristics of the object to detect [VIO 04]. Another piece of information often used to characterize the face is its skin color. In fact, color patterns can be learned to describe the skin color, and associated face detectors can be used for detection [HSU 98]. A review of face detection methods is given in [ZHA 10]. Finally, the reflectance, which is a more physical feature of the face, may also be linked to color. This feature explains the light reflected by a point on the surface and can be related to the perceived intensity in an image in this projected point.

The specific shape of the faces may also be characterized through distances (between the feature points, for example), 2D or 3D point distributions, or surface meshes. The silhouettes, corresponding to the shape border once projected in the image, also provide rich information to estimate the 3D shape of a face.

Both shape and texture of faces can be learned in order to build face models. However, despite the genericity of the class of faces in terms of appearance and shape, it should be noted that there is a large intraclass variability of individuals, which allows one to differentiate an individual from another. It is this difference that should be exploited in identification and authentication algorithms. Some models include both generic aspects and individual properties of faces. These are obtained through a learning process from which an average model (2D or 3D) and deformations are extracted, associated with a probability of occurrence. They characterize either the shape or the texture of the class of faces, or even both jointly.

Model-based approaches have several advantages. First, the use of *prior* information on the shape and/or the texture constrains the space of solutions and allows one to regularize the solution in case of noisy data. Furthermore, the knowledge of an associated model of texture and shape provides rich information for the estimation of the face. In fact, it provides information on areas of interest (feature points, high-gradient areas, silhouette, etc.) and allows one to compute the similarity with the observations made in these areas in order to optimize the parameters.

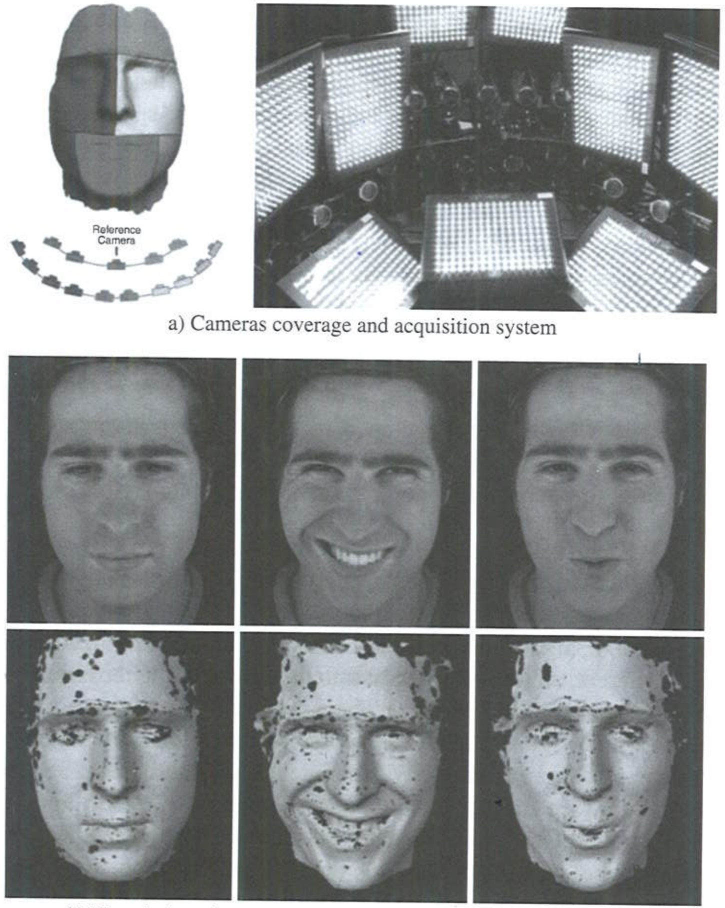
4.3. Geometric approaches for the reconstruction

Many algorithms have been developed to reconstruct an object from a set of images. In the case of face reconstruction, those based on stereovision (or more generally on multiview acquisitions) or a shape from shading are the most used.

4.3.1. *Stereovision – Multiview*

The first type of algorithms are based on a set of synchronized views of the object from different angles and takes into account the stereovision constraints coming from the system’s calibration. The principle is as follows: interest points detected on each view are first matched (possibly constrained by the epipolar lines resulting from the calibration data). We then deduce the associated 3D positions to recover the 3D information of the object. The non-textured points are then reconstructed by interpolation from the sparse set calculated in the previous step, or by using the epipolar constraints again. A detailed description is given in [SEI 06], where the authors categorize the different algorithms according to the initialization, the methodology, and the *prior* information used.

These methods impose several constraints. First, it is essential to have a significant number of corresponding points on the entire surface of the face in order to obtain a valid reconstruction at any point. Therefore, it is necessary to have views taken from close angles to satisfy this condition; otherwise, a point is not necessarily visible in different images. Moreover, the calibration parameters should be accurately known to perform the correct triangulation of matching points. However, some methods estimate the 3D shape of an object from a set of views acquired when the extrinsic calibration parameters are not completely (or partially) known [DAL 09, POL 04]. The procedure then follows structure from motion approach, which is detailed in section 4.6.3.



a) Cameras coverage and acquisition system

b) Three independent examples of reconstructions during a sequence

Figure 4.4. *Multiview reconstruction proposed in [BRA 10]*

The use of multiview acquisitions for the reconstruction of faces has been proposed several times [BRA 10, LIN 10, BEE 10], with various numbers of views and qualities of sensors. With the rise of high-resolution cameras, the reconstructions increasingly rely on high-quality multiview systems, approaching the accuracy obtained with active acquisition methods (laser scanner, projected light). Although they do not require markers, these systems are sometimes restrictive if they involve multiple sensors and a special lighting system (Figure 4.4(a) illustrates the system proposed in [BRA 10]). However, they lead to a very accurate reconstruction of the observed face (Figure 4.4(b)). In fact, matching is performed on mesoscopic details (skin pores, wrinkles) using the high resolution of images, which provides a dense cloud of points (of the order of 8–10 million points for a face) and a very accurate final mesh. Other reconstruction methods have also been proposed from a single high-resolution binocular system [BEE 10], by exploiting the fine details of the face as before. To reduce the system's cost and the execution time, some methods are based on lower resolution images, at the cost of a less accurate quality reconstruction. Lin *et al.* [LIN 10] use five views of the face with a highly variable pose to reconstruct the face by using the bundle adjustment algorithm and dynamic programming. Using information from silhouette and profile views permit to improve the reconstruction, especially in the nose region, but still remains less accurate than the previous methods.

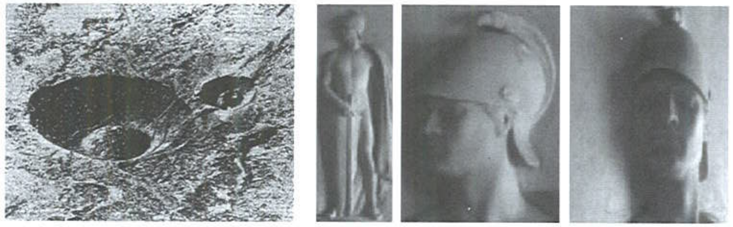
4.3.2. *Shape from shading*

The shape from shading approach [ZHA 99] estimates the geometry of an object from one or more of its views, using the shading. This information characterizes the intensity variation observed in an image, between two points of a surface with identical properties, or of a single point observed in two views with different illumination conditions. As the observed intensity depends on the orientation of the associated surface, shape information of the object may be inferred from shading. This requires not only the model of the optical system, but also the knowledge of the scene illumination and the reflectance properties of the object to be reconstructed. A typical hypothesis for the shape from shading method is to consider the object as Lambertian, meaning that the light reflected from a point on its surface is the same in all directions. Other more realistic models, such as the Phong illumination model, also exist. This takes into account not only the ambient component and the diffuse reflectance (Lambertian model), but also the specular reflection, which characterizes a preferential reflection direction. The intensity *I* of a point is then given by the sum of three terms:

$$I = \underbrace{k_a I_a}_{\text{ambient component}} + \underbrace{k_d I_d \cos\theta}_{\text{diffuse reflection}} + \underbrace{k_s I_d \left(\cos\alpha\right)^{\upsilon}}_{\text{specular reflection}} \qquad [4.1]$$

where I_a and I_d are, respectively, the intensities of the ambient and directional lights, k_a, k_d, and k_s are the ambient, diffuse, and specular reflection coefficients, respectively, θ is the angle between the normal at the considered point and the direction of the directional light, α is the angle between the reflection and viewing directions, and υ is the brightness coefficient of the considered point. This model is more realistic and therefore allows precise shape estimation. Some authors have proposed to specifically measure the reflectance of the face [MAR 99] by learning the *bidirectional reflectance distribution function* (BRDF), which models the light reflection at a point on a surface.

These techniques require *prior* information, such as the position of the light source, for example, when using a single image. Otherwise, the shape from shading problem is ill-posed and it is not possible to directly infer a surface uniquely from an image. Various ambiguities have been shown in the literature, such as crater [PEN 89] and *bas-relief* [BEL 97]. The first is illustrated in Figure 4.5(a) and shows the ambiguity that exists if the lighting was to be jointly estimated with the surface. Here, the lighting can be perceived as coming from above (view of a crater) or from the bottom (view of a volcano upside down): the surface and lighting cannot thus be determined uniquely. Figure 4.5(b) shows an example of ambiguity known as *bas-relief*, where the estimation of the topography of the face estimated by looking at the central image is wrong. In fact, this is actually much more flattened (figure on the right). Different 3D surfaces, combined with suitable light sources, can therefore lead to the same image after projection.



a) Crater. Source: [PEN 89]

b) *Bas-relief*. Source: [BEL 97]

Figure 4.5. *Multiview reconstruction proposed in [BRA 10]*

Initially, the shape from shading method was developed to estimate the shape of an object from multiple views from a fixed pose and different known illumination conditions [WOO 89]. Developments have been proposed to relax these conditions, and not to require the knowledge of the illumination parameters [BAS 07, WU 11]. Recently, a shape from shading method with no pose and light sources constraints has been proposed, making it possible to use a large number of acquisitions to reconstruct the face [KEM 11b]. To reduce the influence of the shape changes in the whole set of images (typically due to the expression variations), the notion of canonical shape, which is defined as the shape locally similar to the largest possible number of photographs, is introduced.

Unlike the stereovision technique that reconstructs an object by interpolation from a sparse set of 3D points, the shape from shading technique estimates the normal vector at each pixel of the image and thus provides more accurate reconstructions. This is particularly the case for low textured surfaces, such as the cheeks of the face, where very few points of interest are detected. Nevertheless, both methods can turn out to be complementary by initializing a solution through multiview stereovision before refining it via shape from shading technique [WU 11].

The advantages of the aforementioned techniques are that they do not use assumptions on the object and thus permit us to reconstruct any object. But faces exhibit very few textured areas, especially on the cheeks or on the forehead. For some of these techniques, it is therefore difficult to infer the 3D shape. Thus, as the shape and the texture of faces can be modeled, it is interesting to use such *prior* information to compute the solution. Modeling the class of faces, on the one hand, allows one to reduce the search space to a suitable subspace and, on the other hand, provides a means to regularize the solution.

4.4. Model-based approaches for reconstruction

4.4.1. *Modeling of the face*

Many studies have been devoted to the modeling of faces in two and three dimensions. We present here a brief overview of the best-known models, and conclude with a more detailed description of the most commonly used 3D model, namely the *3D morphable model* (3DMM) [BLA 99].

The choice of the class of face modeling is constrained by the type of information to be processed (images, depth sensor, inertial system, etc.) and by the application for which the model is used. In fact, for human–computer interaction or video-conference applications, important information is contained in facial movements (expressions, words). A generic model, common to all individuals, is therefore sufficient. However, it is necessary to associate a deformation model related to facial movements, for instance, as for *GRETA* [PAS 01]. On the contrary, when the facial reconstruction is part of a face recognition application, a deformable model is required, where the deformation reflects the specificities of each individual. Let us note that some models combine an identity model and an expression model

[BLA 03a], thus providing greater flexibility, and require more efficient algorithms to estimate all the identity and expression parameters.

4.4.1.1. *2D modeling of the face*

The first face models that appeared in the 1990s were 2D representations. Among others, these include:

Eigenfaces [TUR 91], which are the main vectors extracted from a principal component analysis (PCA) on a database of faces in frontal view. The PCA aims at capturing the variability of the training set and to encode it in a series of vectors in their order of importance. This set of vectors, called *eigenfaces*, defines the basis on which a face can be expressed as their linear combination.

Labeled graphs [WIS 97], which define the face as a labeled graph. Each node of the graph is associated with a vector concatenating responses to a Gabor filter bank around the corresponding point of the face. Each edge is labeled with the distance between the two end points.

Active shape models [COO 95], which statistically characterize the distribution of face shapes (in 2D). The adjustment of the model (in terms of pose and deformation) with an input image is done recursively, by matching the model with the contours or points of interest of the observed image, followed by an update of the pose and shape parameters.

Active appearance models (AAM) [EDW 98], which consider the texture in addition to the statistical shape model. The estimation of the model parameters is done by minimizing the difference between the observed texture in the input image and that generated from the estimated shape and texture.

Most algorithms that estimate the parameters of one of these models, given an image, require a frontal or near-frontal view. Otherwise, these models cannot be fitted to the observed face.

4.4.1.2. *3D modeling of the face*

Given the characteristics of the acquisition system presented in section 4.1.2, it is necessary to manage the images of faces under non-frontal poses. In fact, due to the camera configuration (for instance, on the doorposts, or in a corner of a room), the pose under which the face is perceived can vary considerably. To address this problem, it is common to work with a 3D face model. Thus, the joint estimation of the pose and the model allows one to proceed to frontalization, as defined in Section 4.1.2. In addition, the use of a 3D model is a solution to self-occlusion and shadow issues, if the light sources are integrated into the parameters to be estimated.

A simple 3D face model called *Candide* was proposed in 1987 and consists of a mesh characterizing the frontal part of the head [RYD 87]. This mesh has been modified to match the MPEG4 standard and action units have then been added to characterize expressions (*Candide-3 model* [AHL 01]). However, this model does not characterize the intraclass variability of faces (in terms of identity), resulting in the construction of other models, such as the 3DMM. The seminal paper on 3DMM [BLA 99] proposed by Blanz *et al.* is the source of numerous works on 3D modeling of the face. The main contribution of this paper is the introduction of a statistical model of the face, in terms of shape and texture, from a set of M 3D acquisitions of faces, densely aligned. Each face is described by its shape $\mathbf{S} = \{(X_1, Y_1, Z_1), \ldots , (X_N, Y_N, Z_N)\}$ that consists of N 3D points, and by its texture $\mathbf{T} = \{(R_1, G_1, B_1), \ldots , (R_N, G_N, B_N)\}$. From M faces $\{(\mathbf{S}_i,\mathbf{T}_i), i \in \{1, \ldots , M\}\}$ from which the mean $(\overline{S},\overline{T})$ is removed, the PCA is performed independently on the shape and on the texture, leading to the covariance matrices C_S and C_T. The principal axes of the shape and texture deformation are, respectively, characterized by the eigenvectors $\mathbf{s}_i$ and $\mathbf{t}_i$. A face $(\mathbf{S},\mathbf{T})$ resulting from this modeling is described by:

$$S = \overline{S} + \sum_{i=1}^{M-1} \alpha_i s_i, \quad T = \overline{T} + \sum_{i=1}^{M-1} \beta_i t_i \qquad [4.2]$$

where $\alpha = (\alpha_1, \ldots , \alpha_{M-1})$ is a real-valued vector distributed with a probability:

$$p(\alpha) \approx \exp\left\{ -\frac{1}{2} \sum_{i=1}^{M-1} \left(\frac{\alpha_i}{\sigma_{S,i}} \right)^2 \right\} \qquad [4.3]$$

where $\sigma_{S,i}$ are the eigenvalues of the shape covariance matrix C_S. The probability of the vector of texture coefficients $\beta = (\beta_1, \ldots , \beta_{M-1})$ is expressed similarly. Figure 4.6 shows the influence of the variation of shape parameters α on the overall shape of the face for a given texture. Each face is generated with the same texture, and the projection is applied with the same calibration parameters.

There are two main benefits to define the face by the 3DMM:

– The number of unknowns to be estimated in order to characterize the shape and the texture is greatly reduced. In fact, instead of independently defining thousands of 3D points and their associated color, the PCA reduces the definition of texture and shape to a smaller set of parameters, which measure the eigenvectors.

– The definition of a new face as a combination of eigenvectors selected following the PCA uses a strong *prior* knowledge derived from the training set of faces. Thus, this knowledge allows one to create consistent faces because of their fidelity to the model.

A point to be evaluated with a model built from a training set is its ability to characterize the face of any individual. It may, indeed, not be perfectly reconstructed with the eigenvectors derived from the PCA. Therefore, in this case, we search the parameters $\{(\alpha_i, \beta_i), i = 1, \dots, M-1\}$ such that the distance of the considered face to the face space V defined by the 3DMM is minimum (according to the distance to be defined). The solution is thus the projection of the real face onto V.

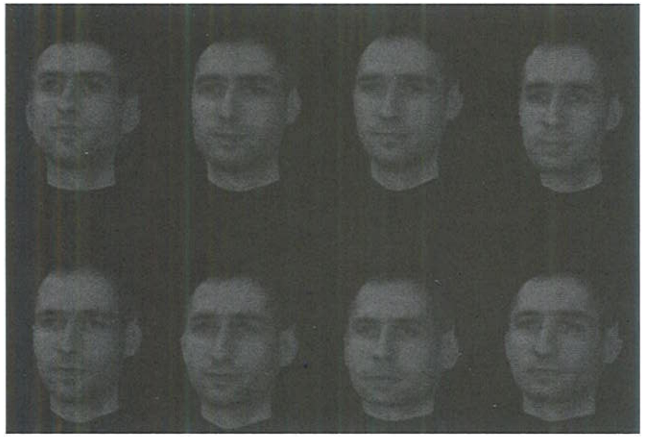


Figure 4.6. *Variation of the projection of a face for different shape parameters, with given pose and texture*

An intermediate between the active appearance model and the 3DMM was proposed by Xiao *et al.* [XIA 04] to characterize faces. However, this model, which characterizes as many shapes as the 3DMM, does not handle the problems of occlusions (since the 3D information is not explicit). Moreover, it is less densely defined than the 3DMM, and can therefore be restrictive for face recognition applications. However, the advantage of this model is its speed of pose and deformation adjustment, given an image, which is similar to that of a conventional AAM, and much larger than the estimation methods of the 3DMM that we will now review.

4.4.2. *Estimation of the model parameters*

In this section, we describe different methods proposed to estimate the parameters $\{(\alpha_i, \beta_i), i = 1, \dots, M-1\}$ (equation [4.2]) of a face observed in one or several images.

4.4.2.1. *Joint shape and texture estimation*

Different criteria can be used to estimate the 3D shape of the face parametrized by the coefficients α_i (equation [4.2]), as well as the associated texture. In [BLA 99], a method is proposed for jointly estimating the face parameters (α, β) and the illumination parameters of the scene as well as the calibration parameters (concatenated in the vector p for clarity). This process is performed by minimizing the overall energy E that consists of a data-fidelity term E_I and a regularization term E_M. The first is expressed by:

$$E_I = \sum_{x,y} \left\| I_{\text{obs}}(x,y) - I_{\text{gen}}(x,y,\alpha,\beta,p) \right\|, \qquad [4.4]$$

where (x, y) characterizes the position of a pixel, $I_{\text{obs}}(x, y)$ is its value in the input image, and $I_{gen}(x, y, \alpha, \beta, p)$ is the one in the generated image, given the current values of the parameters. The regularization term E_M includes the assumption of the normal distribution of the shape and texture parameters:

$$E_M = \sum_{i=1}^{M-1} \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_{i=1}^{M-1} \frac{\beta_i^2}{\sigma_{T,i}^2} \qquad [4.5]$$

The minimization of the energy $E = E_I + E_M$ through stochastic gradient descent is proposed in [BLA 99] in order to be robust to local minima and to increase the execution speed of the algorithm. Romdhani *et al.* [ROM 02] proposed an iterative method for parameter estimation, by exploiting the linearity of the equations when the non-estimated variables are fixed. The method relies on the computation of the optical flow between the synthesized image with the current head estimation and the input image. This algorithm requires knowing the direction of light and the approximate pose to initialize the optimization algorithm. It yields results similar to those of [BLA 99], but with a running time divided by five. However, it does not take into account the shading to estimate the shape, unlike the stochastic gradient. The inverse compositional image alignment method proposed in [ROM 03] follows an adjustment method of the face initially proposed in 2D for AAM in [BAK 01] and is based on the inverse projection of the shape model. To increase the convergence radius of the above methods, Romdhani *et al.* [ROM 05] proposed to increase the number of likelihood criteria when aligning the model to the input image. Thus, the risk of falling into local minima during the optimization procedure is reduced, while increasing the quality of the estimated model. In addition to the data-fidelity term, the *prior* shape, and the texture information, the authors take into account the position of specific face edges and specular reflections in the image. As before, this procedure offers a compromise between the fidelity to the observations (which may be noisy or contain incorrect data as wrong feature point detections) and the *prior* model. In addition, the direction of the light is no longer required as an input, and is also estimated by the algorithm. The additional criteria proposed in the latter algorithm impose some preprocessing steps (edge extraction, generation of distance maps) on the images that are sometimes noisy, and require an accurate weighting of the criteria in their combination.

4.4.2.2. *Shape parameter estimation and texture extraction*

It is possible to limit the face reconstruction to the geometric component of the 3DMM. In fact, since the final aim is to validate the identity of a person, it is important to have a texture as accurate as possible. Instead of deriving it from a learning set, it can be extracted directly from the observations once the pose and shape estimation is performed. Given the variability of individual textures (e.g. skin color and presence of scars), a very large database would be required to ensure that any sample of the population is similar enough to at least one solution in the space defined by the 3DMM. Furthermore, as fewer parameters need to be determined, the computation time for their estimates is reduced. In this section, we therefore focus on algorithms in which only the shape of the model is estimated, with the possibility of extracting the texture from observations in a second step.

By eliminating the texture, it is possible to use less complex criteria than those described in section 4.4.2.1. For instance, it is possible to derive the parameters of a 3D model (for example, the 3DMM) from a set of 2D points detected on the images. The method therefore consists in solving the inverse problem of determining the pose and the parameters α_i, $i = 1, \ldots, M-1$, such that the facial feature points of the model $S = \bar{S} + \sum_{i=1}^{M-1} \alpha_i s_i$ have projections as close as possible to the detected positions.

By using the statistical knowledge of the model, it is possible to establish a cost function composed of two terms [BLA 04, FAG 08]:

– A data-fidelity term, which corresponds to the distance between the 3D point projections, given the estimated model and the detected points.

– A regularization term, derived from the construction of the 3DMM, namely the Mahalanobis norm of the vector of the deformation coefficients (to be compared to equation [4.5]). The latter is weighted by a factor η, which tunes the relative importance of *prior* information with respect to the data fidelity, as illustrated in Figure 4.7.

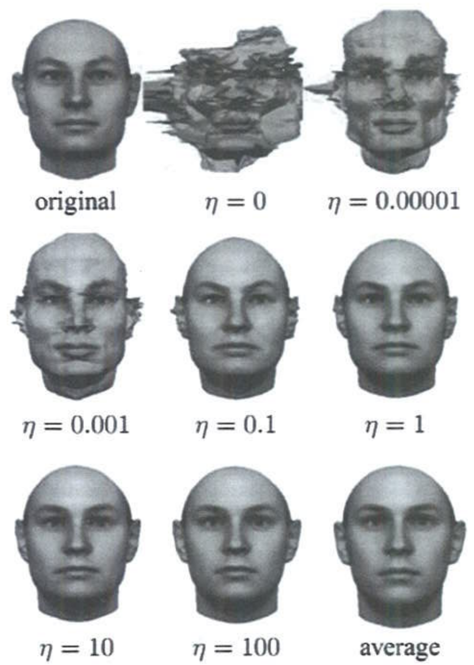


Figure 4.7. *Variation of the reconstruction based on the regularization coefficient η (Source: [BLA 04])*

Thus, the learned model is used to regularize the deformations induced by imprecise detections. In addition to the detected points, it is also possible to add information such as directions tangential to the contours of the face (Figure 4.8).

However, the proposed energy is not robust to outlier detections, which are taken into account in the projection error by using the Euclidean norm. Solutions have been developed to handle detection errors in a better way, by minimizing the constraints of data fidelity. In [BRE 10], for instance, after a few iterations, the energy to be minimized no longer takes into account the 2D distance between the projected points of the model and the detections, but relies on a matching score (through ZNCC, zero-normalized cross correlation) of templates around these points. The more the reconstructed view corresponds to the input image, the better the matching scores and, therefore, the estimated parameters. Thus, the model is not directly adjusted to the detected points but seeks for a compromise between the overall configuration of points of interest and the reconstruction error.

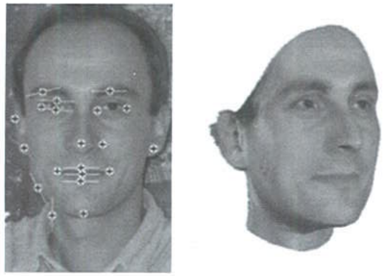


Figure 4.8. *Annotation of the input image and estimation of the associated model (Source: [BLA 04])*

It should be noted that even when points are correctly detected, the feature point positions vary only slightly for faces sampled from the 3DMM. In fact, facial regions with high shape variance are not necessarily located near the main points of the face, and it is, therefore, difficult to capture the deformations through the unique information of feature points. The addition of other criteria, such as the proximity of some edges of the model (lips, eyes) and projected silhouettes with the gradients detected in the image, improves the parameter estimation, and simultaneously increases the complexity and the run time of the algorithm.

A fully automatic method for generating a frontal view from any input image is proposed in [AST 11]. Instead of using the shape part of the 3DMM, the authors used several appearance models, called view active appearance models (VAAM), for different pose intervals. The most appropriate active models are adjusted to the observations, and the one minimizing the residual shape and texture error is kept. An accurate pose estimation is then evaluated for this model with the estimated parameters by support vector machine regression. Finally, the texture of the input image is extracted using an average 3D shape model before generating the corresponding frontal view. The specific shape of each face is not used for the frontalization; the authors refer, indeed, to an average model for this step. Errors can, therefore, exist during the texture extraction when the observed face shape is too different from the average model. However, the time saving is significant compared to a method adjusting to a complete 3D model. Some examples of frontal views generated by this method are shown in Figure 4.9.

Once the pose and the shape parameters are estimated with the methods described here, the texture is extracted, given the 3D shape and a projection model. The whole model (shape and texture) is then used to generate the frontal view. With these methods based on shape adjustment only, the illumination is not considered, and the texture is, therefore, not corrected in case of shadows or specular reflections. Therefore, we have to control the light environment in the acquisition area to minimize its effects. In addition, the extracted texture can be invalid due to 3D

accessories such as glasses. In fact, they are considered to be directly placed against the face during the texture extraction, while they should be spatially modeled in order to separately extract the texture of the glasses on the one hand, and that of the face on the other hand. Otherwise, by changing the pose to get the frontal view, the glass texture will be reprojected onto incorrect areas. One possible solution is to detect the presence of such objects and remove them in the input images (through inpainting algorithms, for example) in order to extract the texture of the face only.

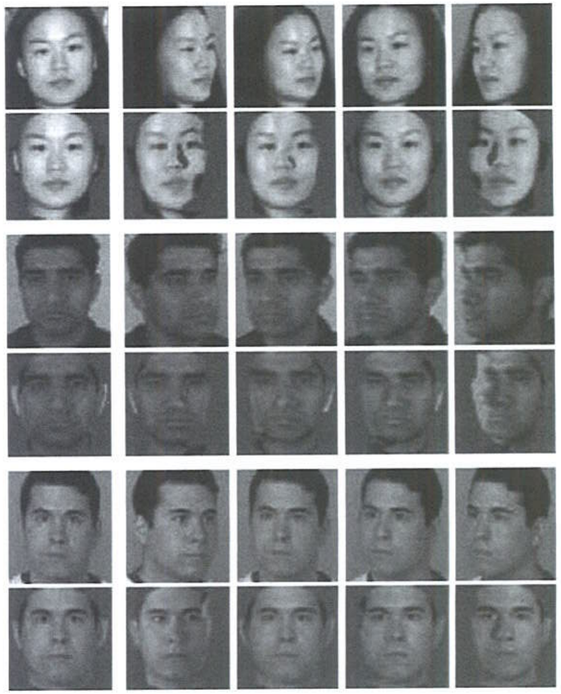


Figure 4.9. *Some examples of faces in frontal view generated from view active appearance models [AST 11]*

One drawback of the model-based approaches is, by construction, their dependence on the training set. Specific attention should be paid to the samples used for the learning step, which should cover all the specificities of faces (e.g. beards and glasses) as much as possible. For example, it is difficult to reconstruct a face with specific scars if this specificity is not present in the training set. Methods that do not rely on a model (section 4.3) are able to reconstruct any facial shapes or accessories worn by the individual, such as glasses, a hat, and a scarf, as they are not constrained by *prior* information. This avoids the problem of inconsistent texture mapping on the face in the case of direct texture extraction.

4.5. Hybrid approaches

The methods described previously (based on the acquisition system properties on the one hand and model-based approaches on the other hand) can be used simultaneously. The problem to be solved thus contains more constraints and input information, which help to solve ambiguities. However, given the amount of information to consider (or estimate), the associated functions are more complex, generally leading to a higher resolution time.

A first possible fusion is to combine the stereovision and the use of an *prior* face model. Given a calibrated system, the shape estimation can be done through the silhouette information extracted in each view [JIN 03, LEE 03], combined in 3D. The advantage of the silhouette feature is its ease of extraction in various poses, not to mention that it is important information in estimating the shape parameters of a face. With multiview acquisitions, the calibration also allows one to compute the 3D positions of feature points such as the eyes and the mouth and to infer the 3D pose as well as the scale of the face. The shape model can then be deformed using points matched between different views [IVA 07]. Finally, the mono-view estimation method of shape and texture presented in [ROM 05] has been extended to the multiview case [AMB 07]. A stereoscopic consistency criterion is integrated into the cost function and the optimization aims at estimating not only the model parameters but also the camera calibration parameters (which do not thus require to be known in contrast to previous methods). From multiple cameras, the depth information of the face is preserved, in contrast to the case of optimization of an image with orthographic projection. The use of multiple views also enables one to solve existing occultation problems when using a unique image and results in more accurate and complete reconstructions (Figure 4.10).

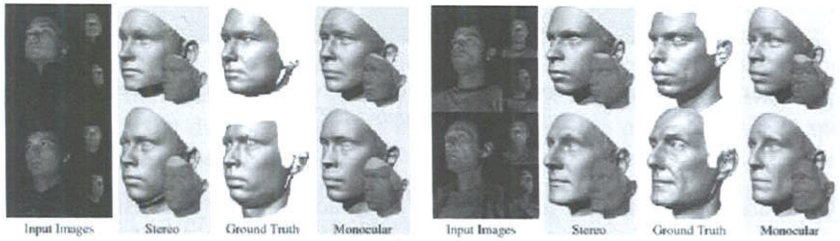


Figure 4.10. *Input images, reconstruction associated with a multiview method, true shape, and reconstruction using only the frontal view. Faces at the bottom of the reconstructions indicate the associated error: the darker the area, the higher the error (Source: [AMB 07])*

Methods combining the shape from shading approach and face models have also been proposed recently. As in [ROM 05], the Levenberg–Macquardt optimization method is used, but the energy to be optimized integrates the constraints derived from the Blinn–Phong reflectance model. The shape parameters of the 3DMM as well as the albedo of the surface can thus be jointly estimated [PAT 09]. To overcome the limitations of a shape model, Kemelmacher-Shlizerman and Basri [KEM 11a] use an average model to estimate the pose and light sources initially. It is then deformed to cope with the observations of a single image, by optimizing the depth of the model points. This method can be advantageous because it does not require learning a deformable shape model, which involves the dense mapping of many 3D acquisitions. Moreover, it helps to reconstruct shapes that are not present in the model.

The main advantage of the methods presented in this section lies in the joint use of *prior* information on the class of objects to be reconstructed, namely the faces, and of photometric and geometric information related to the system. The *prior* assumptions allow fast initialization, while non-model-based methods allow shape and texture reconstruction with a higher accuracy than with the constraints of a model.

4.6. Integration of the time aspect

In sections 4.3–4.5, facial reconstruction is performed from one or more images acquired simultaneously. However, more and more systems now include video sensors, such as the authentication system presented earlier in this chapter (Figure 4.1). It is interesting to exploit the time information to guide the pose estimation (section 4.6.1), and then to multiply the matching scores (section 4.6.2) or even to increase the quality of the head estimation by using frames over the whole sequence (section 4.6.3).

4.6.1. *Face Tracking*

Before estimating the shape and the texture of a face, we should first determine (at least approximately) its position, orientation, and, according to the algorithms, detect some feature points. In fact, most of the algorithms mentioned above are based on these points, and the reconstruction quality depends on the number of detections and their accuracy. In an unconstrained environment, the face is not always seen frontally, which makes the face and its feature point detections a difficult task. Frequent problems are inaccurate points or outliers and non-detections or outliers, or to non-detections. Furthermore, the use of detectors over the entire image is a costly operation, especially if a specific detector is used for each point. If video streams are available, it is interesting to integrate time filtering to guide the face and point detections.

During the last 20 years, many methods have been proposed to track heads in video sequences. Head tracking can be divided into several cases: 2D position (and possibly orientation) tracking using a single camera or 3D position tracking using one or more cameras. An overview of pose estimation methods is proposed in [MUR 09], dealing with pose estimation on a single image and in video streams. In this section, we particularly focus on tracking-based methods, which benefit from the pose history as images arrive recursively.

Several approaches are based on optical flow to recursively estimate the pose of the face, but they are generally constrained by the assumption of constant brightness and require a high frame rate. These methods can be combined with the use of features from an average head model to perform the tracking and minimize these conditions [MAL 00]. Other methods rely on the information provided by face [YAN 06] or feature point [COM 03] detectors to evaluate the pose [ZHU 04]. However, when the object to be tracked presents large appearance changes (in our application, the face pose changes due to its position variation with respect to the sensors), the learning of robust face or point detectors gets difficult. It is then preferable to use approaches that do not rely on detection information.

The Kalman filter [KAL 60], the extended Kalman filter, and the particle filter [DOU 00] are different versions of the Bayesian theory applied to filtering problems. The Kalman filter recursively estimates a state X_t at time t (for example the position of the object of interest in the image I_t) and the associated error in the form of a covariance matrix C_t, given the current observations y_t and the values X_{t-1} and C_{t-1} computed at the previous time $(t-1)$. The Kalman filter relies on Gaussian and linearity assumptions on the functions and noise involved in the process. The particle filter minimizes these conditions based on the approximation of the probability density of the state X_t by a set of particles, each of them representing a hypothesis of the state. Each particle is associated with a weight that characterizes its consistency with observations, and is updated at each new frame. Works based on the particle filtering technique to track the face pose mainly differ by the criteria used to compute the particle likelihood. Conventionally, a color criterion is used to quantify it [PER 02], which is restrictive in the case of illumination and/or pose variations, because the face appearance changes a lot in these cases. In [OKA 05], the particles are estimated from local likelihoods around some feature points, which increases the robustness to pose variations. Kobayashi *et al.* [KOB 06] proposed an original criterion of likelihood by incorporating weak classifiers (based on Haar filter responses) merged by Adaboost into the particle filter. To be robust to pose variations, classifiers can be learned for different intervals of orientation. Thus, the choice of the classifier provides information on the range of the head pose; nevertheless, it does not estimate it accurately. Ba and Odobez proposed to merge the tracking and the pose estimation process [BA 04] in order to estimate both the position and orientation accurately. The particles have a so-called mixed state, characterized by the 2D position of the face in the image on the one hand, and by its orientation on the other hand. To evaluate the particle likelihood, *a prior* learning stage is performed in order to characterize the face responses to Gaussian and Gabor filters for different poses. For a given particle, the observed response in the image is compared to the expected response given its state to evaluate its likelihood. By simultaneously performing the tracking and the pose estimation process, the number of particles to be used to ensure a robust tracking increases exponentially due to the higher size of the search space. The computation time associated with the filtering step is therefore more important, and it simultaneously provides an estimate of the position and orientation of the face and offers robustness to pose changes.

Another way to increase the robustness to illumination and pose changes without a 3D model is updating the appearance features of the object to be tracked [ROS 08, OKA 05]. Nevertheless, by adapting the descriptors to the most recent observations, the update methods potentially suffer from the drift problem, i.e. the insertion of erroneous object features into the appearance model. This bias leads to the accumulation of errors during the tracking and can eventually lead to the lost of the object. To limit this effect, constraints can be imposed on the model update, for example, by controlling the difference between old and new features [KIM 08].

Object tracking under variable poses can also be improved with an explicit 3D model, thus benefiting from the appearance of the object under any pose. This knowledge can be integrated into the particle filter approach, where the likelihood is calculated by comparing the observations with the views that are generated by image synthesis given the particle states [HER 11, BRO 12]. Besides particle filtering methods, optimization algorithms like Gauss–Newton can also be based on a 3D model to perform the tracking, by optimizing the pose parameters explaining the projection of the model on the observed image [MUÑ 09].

The tracking processes detailed in this section can be considered as a preliminary step to shape and texture estimation algorithms. In fact, most of the model-fitting methods require a pose and/or feature point positions as an initialization. The face tracking output provides the first information, which can be used to limit the search regions of point detectors, reducing the processing time for a frame. Besides, the pose filtering allows one to verify the time coherence of successive positions and to detect inconsistent pose values in case of algorithm failure.

4.6.2. *Static approach from video streams*

When video streams are available rather than a single image to recognize a person, the first option is to apply the reconstruction processes presented above to each frame. Thus, there is not only one single comparison that is performed, but also as many as available frames (and therefore frontal views). Even if occlusions are temporarily present, or even if the face is badly estimated at a given frame, others will potentially be valid and therefore usable. To optimize the run time of this method, we must define rules to filter out the bad frames and establish a decision strategy, given all the scores obtained by the matching step between the reference photograph and the reconstructed frontal views.

The frame selection to be processed in order to estimate the face model can be performed through different criteria. The detector confidence, the face resolution, or its pose are commonly used for this purpose [SAT 00, VIL 10]. In addition to these criteria estimated on each image, it is also possible to take into account their time variation between two frames to improve the quality definition. All these criteria should then be merged to define the frame selection process. Besides the usual fusion rules like average, product, and minimum/maximum selection, other methods have also been proposed, such as the use of support vector machines, *k*-nearest neighbors, and the fusion by maximizing the area under the receiver operating characteristics (ROC) curve, which provide the best results according to the comparative study presented in [VIL 10].

Once the frame selection, the face reconstruction, and the frontal view generation are completed, a recognition result for the sequence should be established by fusing the different matching scores. To extend the standard methods of image comparison to video sequences (or a subset of their frames), a distance should be defined between the set of query images (database or passport photo) on the one hand and the video stream on the other hand. One possible definition is the smallest distance computed over all possible pairs (established, for example, in the space of *eigenfaces* [SAT 00]). In the particular case of the identification (where an individual must be selected from a given list), specific criteria can be used to weigh the results obtained on a frame, namely:

– A distance to the model, which characterizes the distance to the nearest individual class. This aims at eliminating the detected faces with a pose or an illumination not represented in the classes.

– A distance to the second closest class to verify the validity of the classification. This criterion is based on the fact that if a class has been selected, it should obtain much better matching scores than the second best class.

By integrating these various criteria into the score fusion process, the classification results are both better than those obtained independently on each frame and those computed through the sum of all scores [STA 07].

The strategies presented in this section exploit the biometric information available in the whole sequence, but consider the reconstruction process independently on each frame. Depending on the number of cameras, on the head pose, and on the algorithms, the facial reconstruction associated with a frame is neither always complete (in case of self-occlusions) nor always precise (noise on the images and insufficient features of the face). The purpose of the next section is to simultaneously exploit the different frames of a sequence to improve the quality of the reconstruction.

4.6.3. *Time consolidation from video streams*

As noted before, the problem of face reconstruction can be ill-posed according to the features used to conduct the pose alignment and the face parameter estimation. The use of multiple views first allows one to disambiguate a point depth derived from its projection in case of a single image. Moreover, as the pose of the face varies over the sequence, occluded regions become visible, thus allowing a complete estimation of the shape and/or the texture. Let us now describe some methods based on the whole video sequences to reinforce the reconstruction.

We can first mention the 2D consolidation methods, which do not require a 3D model. Hu *et al.* [HU 09] proposed the incremental reconstruction of the frontal view of a face by accumulating the regions corresponding to visible areas at each time. A distortion is applied to the observations to align the textures extracted with an average frontal view model. In this way, the areas observed during the sequence lead to a more complete face reconstruction than with a single frame. The advantage of such methods is their speed and independence of complex face models. Nevertheless, the greater the pose angles, the harder it becomes to find the similarity parameters of the deformation.

The use of 3D models provides a greater robustness to pose variations observed in the video streams. A 3D shape model can be estimated from a set of frames in a video, using the silhouette information [SAI 07], detected feature points [FAG 08], or salient points matched between frames of the video [FID 07]. The pose in each frame can be obtained using a specific marker on the face [IAS 07] or can be estimated by a method known as structure from motion, based on a set of matched points of the face, possibly constrained by a generic model [FID 07]. As in the case of some hybrid methods presented in section 4.5, this model can then be abandoned for an accurate reproduction of facial shapes. The structure from motion method typically relies on a matching of specific points between different views, which requires small pose variation between frames to guarantee the visibility of points. The use of an explicit face model creates an intermediate space to link all the observations, and it is then no longer necessary to match the detections between the views, the latter being attached to the model. To estimate accurately the observed shape, it is necessary to have a significant number of points (46 in [FAG 08]) over the entire face. Otherwise, multiple sets of parameters can verify the matching, without a dense validation of the shape similarity (which is similar to some issues identified in section 4.4.2.2). However, this method, which has the advantage of being fast, requires a large number of input feature points that are not always possible to detect due to the head pose in the images. As before, a compromise between the criteria used (related to the reconstruction accuracy) and the execution speed is thus to be found in order to meet the accuracy and speed requirements imposed by the system. A probabilistic approach can also be considered to estimate the shape parameters from a video sequence, for example using the particle filter [HER 12].

An additional difficulty of video sequences compared to the case of synchronized multiview acquisition comes from the variations of the facial appearance between two time instants (e.g. wink and mouth pinch). Some methods use an expression model (like *Candide*) to estimate the facial deformations and derive the expressions [DOR 05, OKA 05, MUÑ 09]. Of course, in order to impose as few constraints as possible on the user, an optimal system should estimate the shape while being robust enough to expression variations. Methods have already been proposed for this purpose [AMB 08], but do not directly exploit images (or video stream). A model of both shape and expression is optimized using geometric information derived from a 3D scanner, and does not exploit the intensity of images nor a time consistency on the expressions. The potential of such methods applied on 2D images and video sequences are a current topic of research.

4.7. Conclusion

Throughout this chapter, we discussed different methods for 3D face reconstruction. Due to the application constraints of a biometric system, we focused on passive methods, relying solely on video acquisitions. The reconstruction is then used to generate the associated frontal view for face recognition purposes. Generic methods such as stereovision and shape from shading can be used, but the integration of *prior* assumptions on head shapes and appearances can improve the quality of the reconstructions, especially within uncontrolled environments (non-frontal poses, varied illumination conditions). Some authors have proposed benefitting from the advantages of both types of methods, and offer very convincing results by mixing a 3D model and stereovision, for example. Finally, by using video, it is possible, on the one hand, to apply time constraints to the estimated poses in order to speed up the initialization and, on the other hand, to benefit from several face estimates, or to improve the face reconstruction. This is especially useful when only one camera is available, to improve the 3D shape estimation quality and to complete the texture during the acquisition. When video streams are used, we should take into account the existence of facial dynamics, related to expressions, movement of the eyes, etc. The robustness of facial reconstruction to expressions in a single image or in a video stream is currently a very active area of research.

4.8. Bibliography

[AHL 01] AHLBERG J., CANDIDE-3 – an updated parameterized face, Report no. LiTHISY-R-2326, Department of Electrical Engineering, Linköping University, January 2001.

[AMB 07] AMBERG B., BLAKE A., FITZGIBBON A.W., ROMDHANI S., VETTER T., “Reconstructing high quality face-surfaces using model based stereo”, *International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 1–8, 2007.

[AMB 08] AMBERG B., KNOTHE R., VETTER T., “Expression invariant 3D face recognition with a morphable model”, *IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, pp. 1–6, 2008.

[AST 11] ASTHANA A., MARKS T.K., JONES M.J., TIEU K.H., ROHITH M., “Fully automatic pose-invariant face recognition via 3D pose normalization”, *International Conference on Computer Vision*, pp. 937–944, Barcelona, Spain, 2011.

[BA 04] BA S., ODOBEZ J., “A probabilistic framework for joint head tracking and pose estimation”, *International Conference on Pattern Recognition*, Cambridge, United Kingdom, vol. 4, pp. 264–267, 2004.

[BAK 01] BAKER S., MATTHEWS I., “Equivalence and efficiency of image alignment algorithms”, *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, pp. 1090–1097, 2001.

[BAS 07] BASRI R., JACOBS D., KEMELMACHER I., “Photometric stereo with general, unknown lighting”, *International Journal of Computer Vision*, vol. 72, pp. 239–257, May 2007.

[BEE 10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M., “High-quality single-shot capture of facial geometry”, *ACM Transactions on Graphics (SIGGRAPH)*, Los Angeles, USA ,vol. 29, no. 4, pp. 40:1–40:9, 2010.

[BEL 97] BELHUMEUR P.N., KRIEGMAN D.J., YUILLE A.L., “The Bas-relief ambiguity”, *IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 1060–1066, 1997.

[BLA 99] BLANZ V., VETTER T., “A morphable model for the synthesis of 3D faces”, *SIGGRAPH*, Los Angeles, USA, pp. 187–194, 1999.

[BLA 03a] BLANZ V., BASSO C., POGGIO T., VETTER T., “Reanimating faces in images and video”, *Computer Graphics Forum, Eurographics*, Thessaloniki, Greece, vol. 22, no. 3, pp. 641–650, 2003.

[BLA 03b] BLANZ V., VETTER T., "Face recognition based on fitting a 3D morphable model", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1063–1074, 2003.

[BLA 04] BLANZ V., MEHL A., VETTER T., SEIDEL H.-P., "A statistical method for robust 3D surface reconstruction from sparse data", *3D Data Processing, Visualization, and Transmission*, pp. 293–300, 2004.

[BRA 10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A., "High resolution passive facial performance capture", *ACM Transactions on Graphics (SIGGRAPH)*, Los Angeles, USA, vol. 29, no. 4, pp. 41:1–41:10, 2010.

[BRE 10] BREUER P., BLANZ V., "Self-adapting feature layers", *European Conference on Computer Vision*, Heraklion, Greece, pp. 299–312, 2010.

[BRO 12] BROWN J.A., CAPSON D.W., "A framework for 3D model-based visual tracking using a GPU-accelerated particle filter", *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, pp. 68–80, 2012.

[COM 03] COMANICIU D., RAMESH V., MEER PP., "Kernel-based object tracking", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[COO 95] COOTES T.F., TAYLOR C.J., COOPER D.H., GRAHAM J., "Active shape models – their training and application", *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[DAL 09] DALALYAN A., KERIVEN R., "L1-penalized robust estimation for a class of inverse problems arising in multiview geometry", *Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 441–449, 2009.

[DOR 05] DORNAIKA F., DAVOINE F., "Simultaneous facial action tracking and expression recognition using a particle filter", *International Conference on Computer Vision*, Beijing, China, pp. 1733–1738, 2005.

[DOU 00] DOUCET A., GODSILL S., ANDRIEU C., "On sequential Monte Carlo sampling methods for Bayesian filtering", *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[EDW 98] EDWARDS G.J., TAYLOR C.J., COOTES T.F., "Interpreting face images using active appearance models", *IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 300–305, 1998.

[FAG 08] FAGGIAN N., ANDREW P.P., JAMIE SHERRAH, "3D morphable model fitting from multiple views", *IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, pp. 1–6, 2008.

[FID 07] FIDALEO D., MEDIONI G., "Model-assisted 3D face reconstruction from video", *International Conference on Analysis and Modeling of Faces and Gestures*, Rio de Janeiro, Brazil, Berlin, Heidelberg, Springer-Verlag, pp. 124–138, 2007.

[HAR 04] HARTLEY R.I., ZISSERMAN A., *Multiple View Geometry in Computer Vision*, 2nd ed., Cambridge University Press, 2004.

[HER 11] HEROLD C., GENTRIC S., MOËNNE LOCCOZ N., "Suivi de la pose 3D du visage en environnement multi-caméras avec un modèle tridimensionnel individualisé", *ORASIS*, Congrès des jeunes chercheurs en vision par ordinateur, Praz-sur-Arly, France, 2011.

[HER 12] HEROLD C., DESPIEGEL V., GENTRIC S., DUBUISSON S., BLOCH I., "Head shape estimation using a particle filter including unknown static parameters", *International Conference on Computer Vision Theory and Applications*, Rome, Italy, pp. 284–293, 2012.

[HU 09] HU C., HARGUESS J., AGGARWAL J.K., "Patch-based face recognition from video", *IEEE International Conference on Image Processing*, Cairo, Egypt, pp. 3285–3288, 2009.

[HUA 11] HUANG H., CHAI J., TONG X., WU H.-T., "Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition", *ACM Transactions on Graphics (SIGGRAPH)*, Vancouver, Canada, vol. 30, no. 4, pp. 74:1–74:10, 2011.

[HSU 98] HSUAN Y.M., AHUJA N., "Detecting human faces in color images", *IEEE International Conference on Image Processing*, Chicago, Illinois, USA, pp. 127–130, 1998.

[IVA 07] IVALDI W., Synthèse de vue frontale et modélisation 3D de visages par vision multi-caméras, PhD Thesis, ISIR/Université Pierre et Marie Curie, Paris 6, 2007.

[JIN 03] JINHO B.M., LEE J., PFISTER H., MACHIRAJU R., "Model-based 3D face capture with shape-from-silhouettes", *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, pp. 20–27, 2003.

[KAL 60] KALMAN R.E., "A new approach to linear filtering and prediction problems", *Transactions of the ASME – Journal of Basic Engineering*, vol. 82, series D, pp. 35–45, 1960.

[KEM 11a] KEMELMACHER-SHLIZERMAN I., BASRI R., "3D face reconstruction from a single image using a single reference face shape", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 394–405, 2011.

[KEM 11b] KEMELMACHER-SHLIZERMAN I., SEITZ S.M., "Face reconstruction in the wild", *International Conference on Computer Vision*, Barcelona, Spain, 2011.

[KIM 08] KIM M., KUMAR S., PAVLOVIC V., ROWLEY H.A., "Face tracking and recognition with visual constraints in real-world videos", *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, 2008.

[KOB 06] KOBAYASHI Y., SUGIMURA D., SATO Y., HIRASAWA K., SUZUKI N., KAGE H., AKIHIRO S., "3D head tracking using the particle filter with cascaded classifiers", *British Machine Vision Conference*, Edinburgh, UK, pp. 1–10, 2006.

[LEE 03] LEE J., MOGHADDAM B., PFISTER H., MACHIRAJU R., "Silhouette-based 3D face shape recovery", *Graphics Interface*, Halifax, Nova Scotia, Canada, pp. 21–30, 2003.

[LIN 10] LIN Y., MEDIONI G.G., CHOI J., "Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours", *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 1490–1497, 2010.

[MAL 00] Malciu M., Prêteux F., "A robust model-based approach for 3D head tracking in video sequences", *IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, pp. 169–174, 2000.

[MAR 99] Marschner S., Westin S., Lafortune E., Torrance K., Greenberg D., "Image-based BRDF measurement including human skin", *Eurographics Workshop on Rendering*, Granada, Spain, pp. 131–144, 1999.

[MAT 09] Matta F., Dugelay J.-L., "Person recognition using facial video information: a state of the art", *Journal of Visual Languages*, vol. 20, pp. 180–187, 2009.

[MOË 10] Moënne-Loccoz N., Roquemaurel B.D., Romdhani S., Gentric S., "Reconstruction à la volée de portraits frontaux par modélisation 3D des visages", *Revue Electronique Francophone d'Informatique Graphique*, vol. 4, pp. 13–19, 2010.

[MUÑ 09] Muñoz E., Buenaposada J.M., Baumela L., "A direct approach for efficiently tracking with 3D morphable models", *International Conference on Computer Vision*, Kyoto, Japan, pp. 1615–1622, 2009.

[MUR 09] Murphy-Chutorian E., Trivedi M.M., "Head pose estimation in computer vision: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 607–626, 2009.

[NEW 10] Newcombe R.A., Davison A.J., "Live dense reconstruction with a single moving camera", *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 1498–1505, 2010.

[OKA 05] Oka K., Sato Y., "Real-time modeling of face deformation for 3D head pose estimation", *IEEE International Conference on Automatic Face and Gesture Recognition*, Beijing, China, pp. 308–320, 2005.

[PAN 03] Pandzic I.S., Forchheimer R., *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, John Wiley & Sons, 2003.

[PAS 01] Pasquariello S., Pelachaud C., "Greta: a simple facial animation engine", *Conference on Soft Computing in Industrial Applications*, Blacksburg, Virginia, USA, 2001.

[PAT 09] Patel A., Smith W.A.P., "Shape-from-shading driven 3D morphable models for illumination insensitive face recognition", *British Machine Vision Conference*, London, UK, 2009.

[PEN 89] Pentland A.P., "Local shading analysis", in Horn B.K.P. (ed.), *Shape from Shading*, MIT Press, pp. 443–487, 1989.

[PER 02] Perez P., Hue C., Vermaak J., Gangnet M., "Color-based probabilistic tracking", *European Conference on Computer Vision*, Copenhagen, Denmark, pp. 661–675, 2002.

[POL 04] Pollefeys M., Van Gool L., Vergauwen M., Verbiest F., Cornelis K., Tops J., Koch R., "Visual modeling with a hand-held camera", *International Journal of Computer Vision*, vol. 59, pp. 207–232, 2004.

[ROM 02] Romdhani S., Blanz V., Vetter T., "Face identification by fitting a 3D morphable model using linear shape and texture error functions", *European Conference on Computer Vision*, Copenhagen, Denmark, pp. 3–19, 2002.

[ROM 03] Romdhani S., Vetter T., "Efficient, robust and accurate fitting of a 3D morphable model", *International Conference on Computer Vision*, Nice, France, pp. 59–66, 2003.

[ROM 05] Romdhani S., Vetter T., "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior", *IEEE Conference on Computer Vision and Pattern Recognition*, Beijing, China, pp. 986–993, 2005.

[ROS 08] Ross D.A., Lim J., Lin R.-S., Yang M.-H., "Incremental learning for robust visual tracking", *International Journal of Computer Vision*, vol. 77, pp. 125–141, 2008.

[RYD 87] Rydfalk M., CANDIDE: a parameterized face, Report no. LiTH-ISY-I-0866, Linköping University, 1987.

[SAI 07] Saito H., Ito Y., Mochimaru M., "Face shape reconstruction from image sequence taken with monocular camera using shape database", *International Conference on Image Analysis and Processing*, Modena, Italy, pp. 165–170, 2007.

[SAT 00] Satoh S., "Comparative evaluation of face sequence matching for content-based video access", *IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, pp. 163–168, 2000.

[SEI 06] Seitz S.M., Curless B., Diebel J., Scharstein D., Szeliski R., "A comparison and evaluation of multi-view stereo reconstruction algorithms", *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, pp. 519–528, 2006.

[STA 07] Stallkamp J., Ekenel H.K., Stiefelhagen R., "Video-based face recognition on real-world data", *International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 1–8, 2007.

[TUR 91] Turk M.A., Pentland A.P., "Face recognition using eigenfaces", *IEEE Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, pp. 586–591, 1991.

[VET 97] Vetter T., "Recognizing faces from a new viewpoint", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, pp. 143–146, 1997.

[VIL 10] Villegas M., Paredes R., "Fusion of qualities for frame selection in video face verification", *International Conference on Pattern Recognition*, Istanbul, Turkey, pp. 1302–1305, 2010.

[VIO 04] Viola P., Jones M.J., "Robust real-time face detection", *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[WIS 97] Wiskott L., Fellous J.-M., Krüger N., Von Der Malsburg C., "Face recognition by elastic bunch graph matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[WOO 89] WOODHAM R.J., "Photometric method for determining surface orientation from multiple images", in HORN B.K.P. (ed.), *Shape from Shading*, MIT Press, pp. 513–531, 1989.

[WU 11] WU C., WILBURN B., MATSUSHITA Y., THEOBALT C., "High-quality shape from multi-view stereo and shading under general illumination", *IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, pp. 969–976, 2011.

[XIA 04] XIAO J., BAKER S., MATTHEWS I., KANADE T., "Real-time combined 2D+3D active appearance models", *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 535–542, 2004.

[YAN 06] YANG T., LI S.Z., PAN Q., LI J., ZHAO C., "Reliable and fast tracking of faces under varying pose", *IEEE International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp. 421–428, 2006.

[ZHA 99] ZHANG R., TSAI P.-S., CRYER J.E., SHAH M., "Shape from shading: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[ZHA 00] ZHANG Z., "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330–1334, 2000.

[ZHA 04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S.M., "Spacetime faces: high resolution capture for modeling and animation", *ACM Transactions on Graphics (SIGGRAPH)*, Los Angeles, USA, vol. 23, no. 3, pp. 548–558, 2004.

[ZHA 10] ZHANG C., ZHANG Z., A survey of recent advances in face detection, Report no. MSR-TR-2010-66, Microsoft Research, 2010.

[ZHU 04] ZHU Z., JI Q., "3D face pose tracking from an uncalibrated monocular camera", *International Conference on Pattern Recognition*, Cambridge, UK, pp. 400–403, 2004.

[ZOL 11] ZOLLHÖFER M., MARTINEK M., GREINER G., STAMMINGER M., SÜSSMUTH J., "Automatic reconstruction of personalized avatars from 3D face scans", *Computer Animation and Virtual Worlds*, vol. 22, pp. 195–202, 2011.

Chapter 5

3D Face Recognition

5.1. Introduction

Three-dimensional (3D) face recognition allows us to deal with some problems related to the pose and lighting conditions. In fact, the 3D information, once obtained through appropriate sensors, is invariant to changes in lighting and pose conditions. Nevertheless, the facial deformation caused by expressions has been one of the challenges that researchers and manufacturers are trying to address. In addition, 3D face recognition requires the 3D acquisition of faces. Not only commercial 3D sensors, but also the solutions proposed by the research community have limitations. These include the range of the sensors, that is 1–2 m, the controlled lighting conditions, the precision, and ultimately the duration of the acquisition.

There are currently two major paradigms of face recognition using the 3D modality: the symmetric recognition where the data in the gallery and the probe data are similar, specifically 3D or 3D + texture, and the asymmetric recognition that uses heterogeneous data from the gallery and from the probe. Thus, the gallery consists of 3D or textured 3D data while the probe data are only texture images or vice versa. The advantage of the latter paradigm is that the use of 3D information is limited. It is also referred to as recognition assisted by 3D.

The structure of this chapter follows to a certain extent the order of steps of the 3D recognition, from the acquisition to the recognition. First, we present the current databases of 3D face recognition. Second, we discuss the 3D acquisition

Chapter written by Mohsen ARDABILIAN, Przemyslaw SZEPTYCKI, Di HUANG and Liming CHEN.