1

Introduction

Face image processing has been a fascinating topic for researchers in computer vision, computer graphics, and multimedia. Human face is so interesting to us partly because of its familiarity. We look at many different faces every day. We rely on our face as a communication channel to convey information that is difficult or impossible to convey in other means. Human faces are so familiar to our visual system that we can easily recognize a person's face in arbitrary lighting conditions and pose variations, detect small appearance changes, and notice subtle expression details. A common question in many researchers' minds is whether a computer vision system can process face images as well as a human vision system.

Face image processing has many applications. In human computer interaction, we would like the computer to be able to understand the emotional state of the user as it enriches interactivity and improves the productivity of the user. In computer surveillance, automatic face recognition is extremely helpful for both the prevention and investigation of criminal activities. In teleconferencing, face image analysis and synthesis techniques are useful for video compression, image quality enhancement, and better presentation of remote participants. In entertainment, face modeling and animation techniques are critical for generating realistic-looking virtual characters. Some video games even allow players to create their personalized face models and put "themselves" in the game.

The last decade has been an exciting period for face image processing with many new techniques being developed and refined. But the literature is scattered, and it has become increasingly difficult for a person, who is new in this field, to search for published papers and to get a good understanding of the state of the art. The goal of this book is to provide a systematic treatment of the technologies that have been developed in this field. We hope that it will serve as a good reference for practitioners, researchers, and students who are interested in this field.

Face image processing is a broad field. It is virtually impossible to cover all the topics in this field in a single book. This book mainly focuses on geometry and appearance modeling, which involves the representation of face geometry and reflectance property and how to recover them from images. Face geometry and appearance modeling is a core component of the face image processing field. In general, any face image processing task has to deal with pose and lighting variations, which can be addressed by applying 3D geometry and appearance modeling techniques. Some of these applications are described in the last part of the book.

In this chapter, we give a literature review and describe the scope of the book and the notation scheme.

1.1 Literature review

Human perception of faces has been a long-standing problem in psychology and neuroscience. Early works can be traced back to Duchenne [49, 50] and Darwin [42]. Duchenne [49, 50] studied the connection between an individual facial muscle contraction and the resulting facial movement by stimulating facial muscles with electrodes to the selected points on the face. Darwin [42] studied the connections between emotions and facial expressions. In 1888, Galton [69] studied how to use iris and finger prints to identify people. In the past decade, there has been a lot of study on what features humans use to recognize faces and objects and how the features are encoded by the nervous system [17, 56, 70, 89].

Early research on computational methods for face recognition can be traced back to Bledsoe [20], Kelly [109], and Kanade [106, 107]. Kanade's thesis [106] is a pioneer work on facial image analysis and feature extraction. In these works, the classification was usually done by using the distances between detected face feature points.

Early 1970s was also the time frame when the first computational method for face image synthesis was developed. In 1972, Parke [162] developed a polygonal head model and generated mouth and eye opening and closing motions. After that, Parke [163] continued his research by collecting facial expression polygon data and interpolating between the polygonal models to generate continuous motion.

In 1977, Ekman and Friesen [52] developed the Facial Action Encoding System (FACS). This scheme breaks down facial movement into Action Units (AUs) where each AU represents an atomic action, which is resulted from the contraction of either a single muscle or a small number of muscles. FACS has been extensively used for both analysis and synthesis tasks.

In early 1980s, as part of the effort to develop a system for performing the actions of American Sign Language, Platt and Badler [170] developed a physically based representation for facial animation. In their representation, face skin deformation was caused by the contraction of muscles, and a mechanism was developed to propagate muscle contractions to the skin surface deformation.

In 1985, Bergeron and Lachapelle [13] produced the famous animated short film *Tony de Peltrie*. They photographed a real person performing 28 different facial expressions. These images were marked manually, and the feature point motions were used to drive their face model. Their work demonstrated the power of facial expression mapping.

In 1986 and 1987, Pearce et al. [165] and Lewis and Parke [124] developed speech-driven animation systems that used phoneme recognition to drive lip movement. In 1987, Waters [233] developed a muscle model for generating 3D facial expressions. Compared to the work of Platt and Badler [170], Waters's model is more flexible in that it avoids hard-coding the mapping from a muscle contraction to the skin surface deformation. This technique was adopted by Pixar in the animated short film *Tin Toy* [115].

In 1988, Magnenat-Thalmann et al. [140] proposed to represent a face animation as a series of abstract muscle action procedures, where each abstract muscle action does not necessarily correspond to an actual muscle.

The 1990s witnessed an exciting period of rapid advancement in both face image analysis and synthesis fields. In 1990, range scanners such as the ones manufactured by Cyberware, Inc., became commercially available. Such scanners were extremely valuable for data collection. In the same year, Williams [239] developed a performance-driven facial animation system. He obtained a realistic 3D face model by using a Cyberware scanner. For tracking purposes, a retroreflective material was put on the live performer's face so that a set of bright spots can be easily detected on the face. The motions of the spots were then mapped to the vertices of the face model.

Terzopoulos and Waters [208, 209, 210] developed a physically based model to not only synthesize but also analyze facial expressions. They used a technique called snakes [108] to track facial feature contours, and estimated the muscle control parameters from the tracked contours. The recovered muscle control parameters can then be used to animate a synthetic face model. Lee et al. [121, 122] developed a system for cleaning up laser-scanned data, registering, and inserting contractile muscles.

Since 1990, there has been a tremendous amount of work on image-based face modeling and animation. In the following, we review the literature along five different but interconnected threads: statistical models and subspace

representation, geometry modeling, appearance modeling, animation, and the modeling of eyes and hair.

1.1.1 Statistical models and subspace representation

In 1987, Sirovich and Kirby [112, 199] proposed to use eigenvectors (called eigenpictures) to represent human faces. In 1991 in their seminar paper [216], Turk and Pentland proposed to use this representation for face recognition, and they call the eigenvectors "eigenfaces." Their work not only generated much excitement in face recognition but also made Principal Component Analysis an extremely popular tool in designing statistic models for image analysis and synthesis. In 1992, Yuille et al. [251] developed a template-based technique for detecting facial features. Their technique can be thought of as an extension to the snakes technique [108] in that the local elastic models in the snakes are replaced by the global deformable templates. In 1995, Cootes et al. [38] proposed to use a statistic shape model, called Active Shape Model (ASM), as a better deformable template. An active shape model is constructed by applying Principal Component Analysis to the geometry vectors of a set of face images. ASM was later extended to the Active Appearance Model (AAM) [37], which consists of a statistic shape model as well as a statistic texture model. In 1997, Vetter and Poggio [220] introduced linear object classes for synthesizing novel views of objects and faces. In 1999, Blanz and Vetter [19] further extended this idea and introduced a morphable model for 3D face reconstruction from a single image. They used the University of South Florida (USF) dataset of 200 Cyberware scanned faces to construct a statistic shape model and a statistic texture model. Given an input face image, the shape and texture model coefficients can be recovered through an analysis-by-synthesis framework. Since then, both the USF dataset and the notion of the linear object classes have become very popular. Linear space representations are described in Chapters 2 and 3. The face-modeling technique of Blanz and Vetter [19] will be described in Section 7.2.

The study of subspace representation of face images under varying illuminations also started to attract much attention in the same period. In 1992, Shashua [194] proposed that, without considering shadows, the images of a Lambertian surface in a fixed pose under varying illuminations live in a three-dimensional linear space. In 1994, Hallinan [82] made a similar observation and proposed to use a low-dimensional representation for image synthesis. In 1995, Murase and Nayar [153] proposed to use low-dimensional appearance manifolds for object recognition under varying illumination conditions.

In 1997, Belhumeur et al. [12] extended Murase and Nayar's appearance manifold method to the case of multiple light sources and shadowing and proposed the illumination cone representation. In 2001, Lee et al. [119] showed that there exists a configuration of nine point light source directions so that the nine images taken under these light sources can be used as the basis for the illumination subspace. In the same year, Ramamoorthi and Hanrahan [177] and Basri and Jacobs [8] independently discovered that nine spherical harmonic basis functions are sufficient to approximate any irradiance map of an object under arbitrary illumination conditions. After that, spherical harmonics became a popular tool for illumination modeling [227, 228, 236, 255, 258]. Spherical harmonics will be described in Section 3.3.

One of the early works on bilinear models was by Brainard and Wandell [24] who tried to isolate the effect of varying illumination from that of varying the surface reflectance. In 1992, D'Zmura [51] proposed to use the bilinear model to recover the surface color (albedo) from images obtained under different illuminations. Around the same period, Tomasi and Kanade [212] published their factorization technique to recover shape and motion from an image sequence. In 1997, Freeman and Tenenbaum [64] proposed techniques to learn bilinear models from training observations and demonstrated how the learned bilinear models can be used for various analysis and synthesis tasks. In 2002, Vasilescu and Terzopoulos [218] proposed a multilinear representation, called TensorFace, to model face images with variations in multiple modes such as identity, expression, illumination, and head pose. In 2005, they [219] introduced a multilinear independent component analysis framework that generalizes the traditional ICA (Independent Component Analysis) to the multilinear case. Bilinear models are described in Sections 2.4.1 and 7.4.

1.1.2 Geometry modeling

In early 1990s, many researchers started to become interested in 3D face reconstruction from images. In 1991, Leclerc and Bobick [117] developed a shape-from-shading technique to reconstruct the 3D face geometry from images where they used stereo processing to initialize the shape-from-shading solver. In 1994, Devernay and Faugeras [46] developed a technique to recover the 3D structure by first computing differential properties from disparity maps, and they applied the technique to face images. In 1996, Proesmans et al. [171] developed a structure-light-based system for acquisition of 3D face geometries. In the same year, Ip and Yin [96] developed a system to construct 3D face model from two orthogonal views.

In 1998, Fua and Miccio [67, 68] developed a system to fit an animated face model to noisy stereo data. Given a video sequence, they treat consecutive frames as stereo pairs or triplets. They assumed that both the intrinsic and extrinsic camera parameters were known a priori. The obtained stereo data was used as input for 3D model fitting. In 1999, Fua [66] extended the system so that it does not require calibration data. Given a video sequence for which the calibration motions are not known, they used a generic face model to regularize the bundle adjustment so that they were able to recover camera motions between successive frames. After that, they generated stereo data and fit a face model as in [67].

In 2000 and 2001, Liu et al. [134, 135, 193, 268] developed a system to construct an animated 3D face model from a video sequence. They used a linear space representation to reduce the number of unknowns in the bundle adjustment process. Their system was so robust that they did many live demos at various events including ACM Multimedia 2000, ACM1:Beyond Cyberspace, CHI 2001, ICCV 2001, and the 20th Anniversary of the PC, where they set up a booth to construct face models for visitors. In 2001, Chen and Medioni developed a stereovision-based face-modeling system using a precalibrated and synchronized camera pair. In 2004, Dimitrijevic et al. [47] developed a structure-from-motion-based system to reconstruct 3D face models from uncalibrated image sequences. Their bundle adjustment technique is similar to [193] in that they also employed a linear space face geometry representation in their bundle adjustment formulation. Unlike [193], they used laser-scanned data (USF dataset) to construct their basis. The system developed by Liu et al. [268] will be described in Section 5.1.

In 2006, Golovinskiy et al. [74] developed an example-based approach to synthesize geometric details. They acquired high-resolution face geometries for people of different ages, genders, and races. The geometric details are extracted and represented with displacement maps. Statistical models of the displacement maps are obtained and used to synthesize plausible geometric details of a new face.

In 2007, Amberg et al. [3] developed a model-based stereo system to recover the 3D face geometry as well as the head poses from two or more images taken simultaneously. Like [193] and [47], they also used linear space geometry representation in their model and motion estimation. But they did not use feature point correspondences. Instead, they used the image intensity difference as the objective function. Note that using image intensity difference would not work well if the head is turning (i.e., using a single camera) due to the illumination changes caused by the head rotation.

1.1.3 Appearance modeling

Another line of work that started at the end of 1990s was the reflectance and illumination recovery of face images. In 1997, Marschner and Greenberg [142] developed an inverse lighting system for face relighting. Given a camera model, the face geometry and albedo, and a set of basis lights, their system produces (either synthesizes or captures) a set of basis images for the face under the basis lights. Given a new photograph, their system searches for a linear combination of the basis images to match the photograph. The linear coefficients are essentially the lighting coefficients. In 1999, Marschner et al. [144] developed an image-based system to measure the Bidirectional Reflectance Distribution Functions (BRDF). They assumed that the surface area of an object is curved and all the points on the measured object have the same BRDF. In 2000, Debevec et al. [43] developed a system to capture spatially varying reflectance properties of a human face's skin area. It requires a 2D array of light sources and a number of synchronized cameras, and the light sources and the cameras need to be calibrated. In 2005, Weyrich et al. [237, 238] developed a system to capture and measure a more general bidirectional surface scattering distribution function, which takes into account the translucent component of the skin reflectance. These systems will be described in Section 6.1.

In 2000, Zhao and Chellappa [270] developed a symmetric shape-from-shading technique to handle illumination variations in face recognition. The estimated shape information was used to generate a prototype image under a canonical lighting condition, and the prototype image is used for face recognition.

In 2001, Nishino et al. [157] developed a technique to recover illumination and reflectance from a small set of uncalibrated images by leveraging specular reflections. This technique will be described in Section 6.3.

In 2003, Wen et al. [236] proposed to use spherical harmonics to reconstruct a pseudo irradiance environment map from a single face image where the face geometry was assumed to be known (a generic face mesh was used in their experiment). The pseudo irradiance environment map can then be used to modify the lighting conditions of the face image. In the same year, Zhang and Samaras [255] proposed to recover the nine spherical harmonic basis images from a set of example images in a bootstrap manner. The technique of Wen et al. will be described in Section 6.2.

A physiologically based skin color and texture synthesis technique was proposed in 2003 by Tsumura et al. [214]. Given an input image of a face, they first remove the shading and then extract the hemoglobin and melanin information using independent component analysis. Based on the extracted hemoglobin and

melanin information, they are able to synthesize the texture changes in pigment caused by aging or application of cosmetics. This technique will be described in Section 9.2.

In 2005, Zhang et al. [258] developed a 3D spherical harmonic basis morphable model. The model was built from a set of example face meshes such as the USF dataset. For each face mesh, nine spherical harmonic basis images were computed. In this way, they obtained a matrix of spherical harmonic basis images, which was used as the basis for a bilinear representation of the space of all the shaded faces under arbitrary illumination conditions. In the same year, Lee et al. [118] developed a bilinear illumination model to reconstruct a shape-specific illumination subspace. It requires a large dataset collected in a controlled environment in order to capture the wide variation of the illumination conditions. In 2007, Wang et al. [227, 228] proposed a spatially varying texture morphable model. They divide the image into multiple subregions and have a separate texture morphable model for each subregion. The spatial coherence between subregions is modeled as a Markov random field. Their technique is capable of handling harsh lighting conditions that cause cast shadows and saturations, as well as partial occlusions. The technique of Lee et al. will be described in Section 7.4. The techniques of Zhang et al. and Wang et al. will be described in Sections 7.3 and 7.5, respectively.

1.1.4 Animation

Toward the end of 1990s, encouraged by the success of image-based rendering work, many researchers investigated example-based techniques to generate photorealistic facial animations. In 1997, Bregler et al. [26] developed a video-rewrite system that synthesizes lip-synchronized face animations from speech. This technique was later extended by Brand [25], Ezzat et al. [53], and Joshi et al. [104]. In 1998, Guenter et al. [80] developed a system with six synchronized cameras to capture 3D facial expressions. In the same year, Pighin et al. [168] used convex combinations of the geometries and textures of a person's example facial expressions to generate new facial expressions for the same person. Basically the geometries and textures of a person's example expressions form the basis of the person's expressions. This representation was also used for facial expression tracking [169]. In 2001, Liu et al. [130] proposed a technique, called the expression ratio image, to map one person's expression details to a different person's face. Noh and Neumann [159] developed the expression cloning technique to map the geometric motions of one person's expression to a different person. In 2004, Sumner and Popovic [204] developed a technique to transfer detailed geometric deformations between two 3D

face meshes. In 2006, Zhang et al. [259] developed a technique to synthesize photorealistic facial expression details from feature point motions. In 2007, Song et al. [200] developed a general framework for facial expression transfer. Their technique works on both 3D meshes and 2D images. Some of the facial animation techniques will be described in Chapter 8. Video-rewrite technique will be described in Section 11.1.

1.1.5 Eyes and hair

1.1.5.1 Eyes

The amount of research work on modeling face organs such as eyes and mouth has been relatively small. In 1994, Sagar et al. [185] proposed an anatomy-based eye model for surgical simulation. In 2002, Lee et al. [120] developed an eye movement model based on eye-tracking data of face-to-face conversations. In 2003, Itti et al. [100] developed a neurobiological model of visual attention including the eye and head movement. Based on eye gaze tracking data, a visual attention model was built to predict the gaze direction for any given visual stimuli (e.g., an image). A model for the eye and head movement was also developed to generate animations for avatars.

1.1.5.2 Hair

In comparison, hair modeling has received a lot more attention. Ward et al. [230] provided a nice literature survey on the hair modeling.

One of the earlier works on furry surface rendering was published in 1989 by Kajiya and Kay [105]. They proposed a volumetric texture representation, called texel, to render furry surfaces. They generated some impressive rendering results of a teddy bear model. In 1991, LeBlanc et al. [116] proposed an explicit hair geometry model by representing an individual hair strand with a 3D cylinder. In this way, they were able to leverage existing rendering pipelines to quickly generate hair images. In the same year, Rosenblum et al. [181] developed a technique that uses a mass-spring system to control the motion of a hair strand. In 1992, Anjyo et al. [99] developed a technique for modeling hair style and dynamic behavior. They derived simple differential equations that coarsely approximate physics while capturing the aesthetic features of human hair. In the same year, Watanabe and Suenaga [232] presented a technique to group the individual hairs into wisps for more efficient rendering and animation. In 1993, Daldegan et al. [39] developed an integrated system that combines hair modeling, rendering, animation, and collision detection. They used the wisp model to reduce the complexity of collision detection.

In 1995, Stam [202] proposed to use motion vector field for hair modeling by tracing particles through the vector field. The trajectory of the particle determines the shape of the hair strand. In 1997, Goldman [73] accelerated Kajiya and Kay's method [105] by using an aggregated lighting model. In 1999, Chen et al. [31] developed a hair rendering system using a trigonal prism wisp model. In 2000, Kong and Nakajima [113] proposed a jittering and pseudo shadow technique to improve the realism of hair rendering. Kim and Neumann [110] developed a technique to simulate hair-hair interactions by using a bounding volume that encloses the hair surface. Hadap and Magnenat-Thalmann [81] used fluid flow to model hairstyle by tracing particle trajectories along the vector field of the fluid flow. Essentially, the vector field of the fluid flow acts as a global controller over the hairstyle.

In 2001, Yu [250] used a 3D vector field for hairstyle modeling. They introduced a set of vector field primitives for both global and local control of hairstyles. In 2002, Chang et al. [29] developed a system to model the interactions of the hair strands by applying dynamics simulation on a sparse set of guide strands, and a dense hair model is obtained by interpolating the sparse guide strands. In the same year, Kim and Neumann [111] developed a multiresolution hair modeling system that allows hairstyle editing. At the coarsest level, the hairs are created with a small set of hair clusters. These clusters can be refined by subdividing the clusters. The system allows editing at any level. Grabli et al. [76] developed an image-based technique to recover the geometry of the hair strands from images taken at a fixed view point under various lighting conditions.

In 2003, Marschner et al. [141] measured light scattering from hair fibers and proposed a hair shading model that is more accurate than the model proposed by Kajiya and Kay [105]. Bertails et al. [58] proposed use of an adaptive wisp tree to allow dynamic splitting and merging of hair clusters over time. Ward and Lin [231] developed a level-of-detail representation to adaptively group and subdivide hair during dynamics simulation.

In 2004, Paris et al. [161] developed an image-based technique to capture the hair geometry by extending the system of [76] to allow multiple view points. In this way, they were able to capture the geometry of the entire hair surface.

In 2005, Wei et al. [234] proposed another image-based approach to capture the hair geometry. Compared to the previous work [76, 161], their system does not require special setup for the controlled illumination. A user can simply use a handheld camera to capture hair images under uncontrolled illumination conditions.

In 2006, Bertails et al. [14] proposed to use Kirchhoff's equations for elastic rods to describe the motion of hair strands. The equations are solved by