EUROGRAPHICS 2018
K. Hildebrandt and C. Theobalt
(Guest Editors)

Volume 37 (2018), Number 2
STAR – State of The Art Report

State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications

M. Zollhöfer[1,2] J. Thies[3] P. Garrido[1,5] D. Bradley[4] T. Beeler[4] P. Pérez[5] M. Stamminger[6] M. Nießner[3] C. Theobalt[1]

[1]Max Planck Institute for Informatics [2]Stanford University [3]Technical University of Munich [4]Disney Research
[5]Technicolor [6]University of Erlangen-Nuremberg

Figure 1: *This state-of-the-art report provides an overview of monocular 3D face reconstruction and tracking, and highlights applications.*

Abstract

The computer graphics and vision communities have dedicated long standing efforts in building computerized tools for reconstructing, tracking, and analyzing human faces based on visual input. Over the past years rapid progress has been made, which led to novel and powerful algorithms that obtain impressive results even in the very challenging case of reconstruction from a single RGB or RGB-D camera. The range of applications is vast and steadily growing as these technologies are further improving in speed, accuracy, and ease of use.

Motivated by this rapid progress, this state-of-the-art report summarizes recent trends in monocular facial performance capture and discusses its applications, which range from performance-based animation to real-time facial reenactment. We focus our discussion on methods where the central task is to recover and track a three dimensional model of the human face using optimization-based reconstruction algorithms. We provide an in-depth overview of the underlying concepts of real-world image formation, and we discuss common assumptions and simplifications that make these algorithms practical. In addition, we extensively cover the priors that are used to better constrain the under-constrained monocular reconstruction problem, and discuss the optimization techniques that are employed to recover dense, photo-geometric 3D face models from monocular 2D data. Finally, we discuss a variety of use cases for the reviewed algorithms in the context of motion capture, facial animation, as well as image and video editing.

CCS Concepts

•***Computing methodologies*** → ***Reconstruction; Tracking;*** *Motion capture; Shape modeling; 3D imaging;*

1. Introduction

Human faces occupy a very central place in human visual perception since they play a key role in conveying identity, message, emotion, and intent. For these reasons, the computer graphics and vision communities started very early in building computerized tools for analyzing real-world faces with the goal of generating digital face images [Wil90, GGW*98], a process generally referred to as face capture. The range of application domains that can benefit from these face capture technologies is vast and steadily growing as these methods are improving in speed, accuracy, and ease of use.

In some contexts, detailed face analysis aims at helping a system take specific actions or decisions. These include facial biometrics, face-based interfaces, visual speech recognition, and face-based search in visual assets. In other applications, such as creating

personalized avatars or 3D printing of faces for entertainment or medicine, an accurate three-dimensional reconstruction of a specific face's geometry is of paramount importance. Finally, in a variety of settings, capturing the deformable geometry of a face as well as its texture is the foundation for generating photo-realistic faces and modifying the appearance/performance of a face in video content. Such technologies are required in a wide range of applications, such as the modification (touch-up, editing, completion, reenactment) of real faces for professional visual effects (VFX), the generation of completely digital photo-realistic faces in high-end productions (movies, commercials, music videos, computer games), and for facial augmented reality (AR) and virtual reality (VR) in entertainment, social media and communication.

While the extraction of sparse 2D information such as facial landmarks might suffice in some of these scenarios, the estimation of a dense 3D photo-geometric model (i.e., recovering detailed models of geometry and appearance) of a face from visual data is the holy grail and a key enabling technology. On the capture side, various setups can be employed depending on the target application and on the available resources. These setups range from expensive professional large multi-view stages, possibly with controlled lights and additional depth cameras, all the way to lightweight capture with a mobile phone RGB camera, with or without depth sensing. Latter setups have witnessed remarkable progress in recent years fueled by the consumers' appetite for face technologies and the constant progress in the development of commodity sensors (consider, for example, the recent iPhone X released by Apple with *Face ID* unlocking technology). This increasing demand for lightweight capture solutions led to the development of new high-quality priors that better constrain the challenging and ill-posed monocular reconstruction problem. It also led to a rapid development of novel and powerful face reconstruction and tracking algorithms that perform well on monocular RGB and RGB-D input, even at real-time frame rates. At the heart of such technologies lies the ability to render detailed deformable face models and to inversely estimate their parameters from real images. Working with unconstrained monocular content recorded in general real world scenes not only opens up many consumer applications, but is also of major interest for professional face related tasks in current content creation and animation pipelines, which are still highly manual, complex and time consuming.

Face performance capture has been advanced from two different directions. At one end of the spectrum are off-line approaches that aim for the highest possible quality, often based on complicated and controlled multi-view setups. On the other end are lightweight setups that work with a single monocular RGB or RGB-D camera and nowadays even achieve real-time performance by trading reconstruction quality for increased speed. Current off-line high-quality approaches are getting faster and work in less constrained settings, while recent real-time techniques obtain higher and higher reconstruction quality based on the availability of novel high-quality face priors.

Motivated by this progress, this state-of-the-art report focuses on the developments that have narrowed down the gap between high-end and commodity setups. More specifically, we show how recent methodical and algorithmic progress has helped to achieve remarkable success using only lightweight capture setups that consist of either a single monocular RGB or RGB-D camera. In addition to the algorithmic development, we thoroughly discuss open challenges, unresolved limitations, and possible avenues for future research to further advance the field. Our discussion focuses on methods in which optimization-based reconstruction is the central concept to recover and track a three dimensional model of the face, see Fig. 1. We will review in detail:

1. The lightweight RGB and RGB-D camera systems that are employed in current state-of-the-art methods,
2. the concepts of image formation as well as the simplifications and assumptions of current state-of-the-art approaches that are employed to make monocular reconstruction feasible,
3. the priors that are currently used to better constrain the underlying ill-posed reconstruction problems and how such models can be acquired based on high-quality capture setups,
4. the algorithms to extract accurate 3D deformable, textured face models from monocular RGB and RGB-D data based on the described priors at off-line and real-time frame rates,
5. and how the obtained reconstructions of geometry, texture, and illumination can be used within applications to accomplish complex face editing tasks from performance-based animation to real-time facial reenactment.

2. Related Surveys and Course Notes

In the fields of computer vision and computer graphics, extensive research exists in the areas of facial capture, tracking, animation and recognition. While this survey focuses primarily on methods for monocular 3D facial capture, we direct interested readers to additional surveys and course notes in related topic areas. Perhaps most closely related is the image-based 3D face reconstruction survey of Sylianou et al. [SL09], which covers early 3D facial acquisition approaches from single and multiple viewpoints. Alongside geometry reconstruction, appearance capture is also of great importance. A recent report by Klehm et al. [KRP*15] outlines related methods for capturing the appearance of faces, with the goal of re-rendering in CG. Facial capture is closely related to the problem of non-rigid surface registration, where the goal is to align a specific surface to images or 3D scans, a topic that is covered in detail in recent tutorials [BTP14, BTLP16]. In the context of computer animation, excellent surveys exist that discuss facial rigging [OBP*12] and blendshape based facial animation [LAR*14]. Finally, driven by applications in security and person identification, there is a large body of work on face recognition [ZCPR03, ANRS07], and more specifically expression recognition [SZPY12], which aims to advance user experiences in human-computer interaction scenarios.

3. Overview of the State-of-the-art Methods

Lightweight face tracking and reconstruction has received an impressive amount of attention in the computer vision and computer graphics literature over the past few years. Tab. 1 gives an overview of current monocular state-of-the-art 3D facial performance capture approaches that do not require facial markers and are based on a single commodity RGB or RGB-D camera. We focus our discussion on methods that rely on optimization-based reconstruction to

	Method	Vlasic et al. [VBPP05]	Weise et al. [WBLP11]	Chen et al. [CWS*13]	Li et al. [LYYB13]	Garrido et al. [GVWT13]	Cao et al. [CWLZ13]	Bouaziz et al. [BWP13]	Cao et al. [CHZ14]	Shi et al. [SWTC14]	Suwajanakorn et al. [SSS14]	Thies et al. [TZN*15]	Cao et al. [CBZB15]	Hsieh et al. [HMYL15]	Saito et al. [SLL16]	Thies et al. [TZS*16a]	Garrido et al. [GZC*16]	Thies et al. [TZS*16b]	Wang et al. [WSXC16]	Wu et al. [WBGB16]
Input	RGB	✓	-	-	-	✓	✓	-	✓	✓	✓	-	✓	-	✓	✓	✓	-	✓	✓
	RGB-D	-	✓	✓	✓	-	-	✓	-	-	-	✓	-	✓	-	-	-	✓	-	-
Prior	Graph	-	-	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
	Linear Model	✓	✓	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Anatomical	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
Tracker	Features	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Dense Geometry	-	✓	✓	✓	-	-	✓	-	-	-	✓	-	✓	-	-	-	✓	-	-
	Dense Color	-	-	-	-	✓	-	-	-	-	✓	✓	✓	-	-	✓	✓	✓	-	✓
Output	Geometry	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Details	-	-	-	-	✓	-	-	-	✓	✓	-	✓	-	-	-	✓	-	-	✓
	Reflectance	-	-	-	-	✓	-	-	-	✓	✓	✓	-	-	-	✓	✓	✓	-	-
	Occlusion	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-
	Controllable Rig	✓	✓	-	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speed	Offline	✓	-	✓	-	✓	-	-	-	✓	✓	-	-	-	-	-	✓	-	-	✓
	Online	-	✓	-	✓	-	✓	✓	✓	-	-	✓	✓	✓	✓	✓	-	✓	✓	-
	GPU	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	✓	-	✓	-	-

Table 1: *Overview of current optimization-based monocular state-of-the-art 3D facial performance capture approaches that do not require facial markers and are based on commodity RGB or RGB-D setups. We discuss the foundations of these approaches in more detail. Our discussion focuses on their differences in terms of the used input data, the underlying priors, the employed tracking constraints, the computed output channels and the achieved runtime performance.*

recover and track a three dimensional model of the face. We also briefly discuss how these approaches differ from and compare to recent machine learning-based face capture methods. The current state-of-the-art differs along multiple orthogonal dimensions (see Tab. 1), such as the input channels used, the underlying face prior, the constraints used during tracking, the channels of the computed output reconstruction and the achieved runtime performance. In the following, we will discuss each of these dimensions in more detail.

4. Input Modalities

Facial performance capture techniques commonly aim to reconstruct high-quality 3D face models with detailed facial motion and appearance from optical sensor measurements of a subject's performance. In the past, state-of-the-art approaches have achieved this goal through complex indoor capture setups with dense camera arrays and complex lighting setups that are expensive to build and operate. Recent methods have pushed the frontier of capture further using low-cost and standard monocular devices. In the following, we provide an overview of the different input modalities, from a brief review of multi-view capture setups to monocular RGB and RGB-D sensors, which are the main focus of this report.

4.1. Multi-View Setups

Multi-view capture setups employ calibrated dense camera arrays with controlled indoor illumination [HZW*04, ZSCS04, PL06, BBA*07, FP09, BHPS10, BHB*11, FJA*14]. To aid 3D reconstruction, some setups also resort to marker data [HZW*04, BBA*07, ARL*10, HCTW11, AFB*13] and/or use active illumination [WHL*04, GFT*11, MHP*07, WGP*10, KH12], i. e., structured light patterns cast by a projector. Similar setups are already commercially [VIC, MOV, DI4, IM] available and can be used for high end production. Standard multi-view setups consist of a set of pairwise stereo cameras distributed around the subject's face. From each stereo pair, the visible face geometry is reconstructed via triangulation. The individual reconstructions are then aggregated while enforcing global consistency. While not the focus of this report, such high-quality multi-view techniques have been extremely successful in the past and have enabled the creation of the first photo-real digital actors [BL03, ARL*10, AFB*13, SEL17]. These tech-

niques set a high baseline on which more lightweight monocular approaches can be evaluated.

4.2. Monocular RGB

An RGB camera captures three channel images that separately encode the amount of received red, green and blue light. High-end cameras often use three independent CCD sensors to acquire the three color signals. Many consumer-grade cameras reconstruct the final RGB image from a Bayer pattern, interleaving colored filters in front of a single sensor. Consumer-grade RGB cameras are ubiquitous and are utilized for a wide range of purposes: television production, surveillance and/or security, live streaming, and entertainment, among others. Since RGB cameras are available to most users, researchers usually prefer this input to develop offline [GVWT13, GZC*16, IBP15, SWTC14, SSS14, WBGB16] and online [CWLZ13, CHZ14, CBZB15, SLL16, TZS*16a] monocular reconstruction and tracking approaches. Monocular face reconstruction and tracking is a highly challenging and ill-posed problem, since the image formation process convolves multiple physical dimensions (geometry, surface reflectance and illumination) in a single color measurement. Therefore, current state-of-the-art approaches rely heavily on several simplifications of real world image formation (see Sec. 5) and employ data-driven priors (see Sec. 6).

4.3. Monocular RGB-D

Unlike RGB cameras, commodity RGB-D sensors capture both color and depth data at real-time rates. This helps in resolving the inherent depth ambiguity of the monocular reconstruction problem, since at least a coarse geometry estimate is available. Therefore, such cameras are frequently employed in state-of-the-art monocular face reconstruction and tracking approaches [WBLP11, LYYB13, BWP13, HMYL15, TZN*15]. Existing consumer depth sensors can be classified in either being active or passive devices. Most commonly passive depth sensors are implemented via a stereo camera setup. If a point is found in both views of a calibrated stereo setup, the 3D point can be reconstructed by triangulation. Finding a corresponding point for one pixel of the first image in the second view, is a hard problem. By utilizing the epipolar geometry of calibrated cameras, the search problem can be reduced to a 1D search. The search is based on features (e.g., color, color gradients, edges), but if there are no unique features the search fails. This problem occurs for example if a white wall should be reconstructed. Therefore, one of the cameras in the stereo setup can be replaced by a projector (which can be assumed as an inverse camera). Using a known structured pattern that is projected, correspondence search is simplified. These structured light cameras are widely spread (e.g., Micosoft Kinect, Primesense Carmine, Intel Realsense) and give relatively good depth in the near range, which is important for face tracking tasks. One common problem of structured light sensors is that they cannot reconstruct geometric features and structures that are smaller than the projected pattern. In addition, the captured geometry is slightly smaller at object silhouettes (depth discontinuities) than the real object, since the pattern cannot be reliably identified at such locations. To avoid that humans can see this pattern, these devices typically work in the infrared (IR) domain. Active depth cameras that work in the IR domain have problems under strong sun light, since the sun's own IR radiation overpowers the projector. This problem is tackled with cameras like the Intel Realsense R200 that combine active and passive stereo using two IR cameras and one projector. In an outdoor environment the projector has no effect over the IR light of the sun and the system works as a classical passive stereo setup.

Another type of active depth cameras are time-of-flight (ToF) cameras. These RGB-D cameras compute depth by measuring the round-trip time of a light pulse (e.g., Creative Senz3D or Micosoft Kinect One). Accurately measuring the round-trip time is a big challenge due to the speed of light. Some devices measure the actual round-trip time with rapid shutters, while others indirectly measure it with time-modulated light. In general, these cameras have a poor depth resolution and a lower signal to noise ratio in the near range than their structured light counterparts. They also suffer from noise and systematic distortions that are typical to the measurement principles, such as multi-path errors. Another common problem are 'flying pixels' at depth discontinuities, which have a depth value somewhere between the foreground and background object. These mixed depth values occur, since light from both objects is reflected and influences the depth value computed for the pixel. These can be hard to model and compensate for, which may complicate face capture. Since the near capture range is typically the most important range for face tracking, most state-of-the-art approaches rely on structured light sensors.

5. Image Formation Models

Many of the described methods interpret the face reconstruction problem as an inverse rendering problem. This requires them to mathematically model the image formation process. The geometry of a face is commonly described with triangle or quad meshes. In most cases, also material properties describing how the skin or other elements of the face interact with light, have to be modeled. Modeling the light surface interactions for faces is highly complex, since faces exhibit a certain amount of diffuse reflection and different amounts of specular reflection (with different lobes). Specular skin reflection depends on several factors, such as the oiliness of the skin or sweat. In addition, faces also exhibit a certain amount of transmission and subsurface scattering. An in-depth discussion of skin appearance modeling is beyond the scope of this state-of-the-art report and we would like to refer to [KRP*15] for more details. In the context of monoculuar face capture, most work relies on simple appearance models that approximate real face appearance. Often faces are assumed to be diffuse, modulated by a simple reflectance map. This is a stark oversimplification of reality, which ignores the specular lobe and subsurface effects [JMLH01, BL03], but makes the inverse reconstruction problem easier to tackle. Another simplification is that mostly directional illumination or local low-frequency illumination is assumed to make the problem tractable (e.g., Spherical Harmonics). In Appendix B, we discuss the most commonly employed models. Using these material properties, the geometry and an illumination model, the amount of light on a surface point that is reflected to the camera can be computed.

The mapping from the 3D geometry to a 2D image is established through projective camera models. Given a 3D world point $\mathbf{v}$, a

2D image point $\mathbf{p}$ is obtained using the following general model [FP12]:

$$\mathbf{p} = \mathbf{K}\Pi(\mathbf{R}\mathbf{v}+\mathbf{t}) = \mathbf{K}\Pi(\hat{\mathbf{v}}) \ , \tag{1}$$

where $[\mathbf{R}|\mathbf{t}] \in \mathbb{R}^{3\times4}$ refers to the camera extrinsics that transform the 3D point $\mathbf{v}$ into a point $\hat{\mathbf{v}}$ in camera space. $\Pi(\cdot)$ denotes a potentially non-linear operator that projects the aligned 3D point $\hat{\mathbf{v}}$ onto the 2D image plane, and $\mathbf{K}$ is the geometric property of the camera a. k. a. camera intrinsics. Here, $\mathbf{p} = [\mathbf{p}_x, \mathbf{p}_y, 1]^\top$ is the projection of $\mathbf{v}$ onto the image plane in homogeneous coordinates. In non-homogeneous screen space, we can represent this point as $\hat{\mathbf{p}} = [\hat{\mathbf{p}}_x, \hat{\mathbf{p}}_y]^\top$. Appendix A details the different camera models used in the literature. For more details and topics such as lens distortion correction we refer to [SRT*11].

6. Face Models and Statistical Priors

To allow the reconstruction of a face in the ill-posed monocular setting or based on incomplete and noisy depth data, priors that model the structure and expression of faces are commonly employed. These priors are typically based on the low dimensional subspaces we describe in this section.

6.1. Blendshape Expression Model

Blendshapes are a set of 3D-models of a face, each with a particular expression [LAR*14]. The topology of all models is identical, so the blendshapes only affect the vertex positions. Animation can be achieved by blending between the neutral face and a particular expression, or by blending between different expressions (see Fig. 2 (b) and Sec. 6.1.3).

Blendshapes are extensively used in 3D modeling and animation due to their intuitive representation. In the animation industry, artists typically directly manipulate blendshapes, each of which controls an elementary face expression or a localized face region. Sometimes they work on sparser higher level controls that directly activate a combination of blendshapes. Blendshapes can be obtained from user-specific expressions [WBLP11, WBGB16], generated via statistical analysis from a large facial expression database [CWZ*14, VBPP05], or can be hand-crafted by animation artists [BWP13]. Artistically created blendshape rigs are typically relatively sparse and controlled through semantically meaningful parameter dimensions.

Absolute Blendshape models (Sec. 6.1.1) describe the expressions of a particular face, but as coarse expressions generalize well across individuals, most approaches employ more generic Delta Blendshapes (Sec. 6.1.2) to track and reconstruct deformations of arbitrary faces. Blendshapes can also be automatically transferred to new character rigs with different topology [SP04, LWP10], however such approaches are beyond the scope of this report.

6.1.1. Absolute Blendshape Model

Let the $\mathbf{b}_i^e \in \mathbb{R}^{3n}$ be the m_e+1 blendshapes of a particular face, where $\mathbf{b}_0^e$ is usually the neutral face. A new facial expression $\mathbf{e}$ is obtained by linear combination, yielding the absolute blendshape model:

$$\mathbf{e} = \sum_{i=0}^{m_e} \delta_i \mathbf{b}_i^e \ , \tag{2}$$

where $0 \leq \delta_i \leq 1, \forall i \in \{0,\dots,m_e\}$ denote the blendshape weights. The parameter limits are under the assumption that each individual blendshape models the maximal allowed activation of the corresponding expression. Note, often a soft constraint, instead of a hard constraint, is employed to keep the weights in a reasonable range. To avoid an undesired global scaling factor, usually a convex combination is used, i. e., $\sum_i \delta_i = 1$ [LAR*14].

6.1.2. Delta Blendshape Model

The model in the previous section only describes animations of one particular face. A more convenient and popular representation used by modeling packages (e. g., Maya) and many approaches in the literature [BWP13, GZC*16, IBP15, LYYB13, TZS*16a, WBLP11] models blendshapes as delta variations that linearly add up on top of the neutral face:

$$\mathbf{e} = \mathbf{b}_0^e + \sum_{i=1}^{m_e} \delta_i(\mathbf{b}_i^e - \mathbf{b}_0^e) = \mathbf{b}_0^e + \sum_{i=1}^{m_e} \delta_i \mathbf{d}_i^e \quad \text{(expression)} \ , \tag{3}$$

where $0 \leq \delta_i \leq 1$ and $\mathbf{d}_i^e$, $\forall i \in \{1,\dots,m_e\}$ are the per-vertex 3D displacements w. r. t. the neutral face $\mathbf{b}_0^e$.

These delta vectors generalize well to other neutral faces, i. e., other identities, with same topology. If we combine them with the parametric model from Eqn. 5, we achieve a general face model:

$$\mathbf{s} \ = \ \mathbf{a}_s + \sum_{i=1}^{m_s} \alpha_i \mathbf{b}_i^s + \sum_{i=1}^{m_e} \delta_i \mathbf{d}_i^e \quad \text{(shape + expression)} \ , \tag{4}$$

where the α_i describe the identity and the δ_i the expression.

6.1.3. Blendshape Weights

Although the box constraints imposed on the linear weights δ_i can control the influence of the blendshapes ($\delta_i = 0$ deactivated blendshape; $\delta_i = 1$ fully-activated blendshape), some blendshapes simply cannot be combined together due to their linear dependency (see Fig. 2 (c)). For instance, jointly moving the mouth to the left and to the right can cause face distortions. The combination of semantically similar expressions, e. g., a wide open smile combined with a mouth open, also adds a double effect, resulting in unrealistic deformations. This problem can be alleviated by utilizing pairwise activation constraints of the form $\delta_i\delta_j = 0, \ i \neq j$ for some of the weights [LAR*14], which means that only one of the two blendshapes should be activated. An alternative is to employ a strong prior that enforces sparsity [BWP13] or restricts the activation of the linear weights [LYYB13, TZN*15].

6.2. Parametric Face Models

The most commonly used prior is based on the Morphable Model of Blanz and Vetter [BV99] which learned a low-dimensional face subspace from high resolution laser scans. Alternative parametric face models were built later, using inputs ranging from commodity [CFKP04] 3D face scans to a large number of high-quality 3D face scans [BRZ*16] . Reconstruction of a linear face model from

monocular face images downloaded from the internet was also investigated [KS13].

The model of Blanz and Vetter is constructed from a database of 200 human faces, with neutral expression, that were digitized using a laser scanner (for details see [BV99]). In addition to geometry they also captured the illumination-corrected textures of the faces. Based on non-rigid template-fitting these scans are first registered and aligned in a common coordinate system. The resulting faces share the same topology, but differ in geometry and skin reflectance. The template mesh is a simple triangle mesh consisting of n vertices. To reduce the dimensionality of the dataset, principle component analysis (PCA) is independently applied to geometry and skin reflectance. PCA computes the principle components of a dataset and the corresponding standard deviations based on the assumption of a multi-variate Gaussian distribution of the samples.

Let the $\mathbf{b}_i^s \in \mathbb{R}^{3n}$ and $\mathbf{b}_i^r \in \mathbb{R}^{3n}$ be the m_s shape and m_r reflectance basis vectors, respectively. The vectors $\mathbf{b}_i^s$ contain the stacked x,y,z-components of all vertices and $\mathbf{b}_i^r$ the r,g,b-components respectively. Using this PCA model, new faces (shape $\mathbf{s}$ and skin reflectance $\mathbf{r}$) are synthesized via a linear combination:

$$\mathbf{s} = \mathbf{a}_s + \sum_{i=1}^{m_s} \alpha_i \mathbf{b}_i^s \quad \text{(shape)}, \tag{5}$$

$$\mathbf{r} = \mathbf{a}_r + \sum_{i=1}^{m_r} \beta_i \mathbf{b}_i^r \quad \text{(reflectance)} \ . \tag{6}$$

Here, $\mathbf{a}_s \in \mathbb{R}^{3n}$ is the average face shape and $\mathbf{a}_r \in \mathbb{R}^{3n}$ the average reflectance. New shapes $\mathbf{s}$ and reflectances $\mathbf{r}$ are generated by adding a linear combination of the basis vectors $\mathbf{b}_i^s$ and $\mathbf{b}_i^r$ using weights α_i and β_i, respectively. The corresponding standard deviations $\sigma_s \in \mathbb{R}^{m_s}$ and $\sigma_r \in \mathbb{R}^{m_r}$ are stored in vectorized form. The standard deviation of the shape and reflectance dimension exhibit a rapid falloff. This can be exploited to reduce dimensionality, i.e., instead of using all principle components of the dataset, a lower number can be used without loss of accuracy, e.g., the $m_s = m_r = 80$ vectors used in Face2Face [TZS*16a] explain 86% of the original variance.

(a) (b) (c)

Figure 2: *Blendshape expression model: (a) Neutral face. (b) Expressions obtained by activating different blendshapes. (c) Artifacts due to the linear dependency between certain blendshapes.*

7. Advanced Face Models and Personalized Rigs

The commonly used statistical face model described in the previous section not only regularizes the ill-posed reconstruction problem, but also enables reconstruction results of high quality. Yet, it is still restricted to comparably coarse scale, i. e., it lacks person-specific idiosyncrasies and fine-scale skin detail, e. g., wrinkles. In addition, it disregards anatomical and physical plausibility in the reconstructions. For example, the state-of-the-art Face2Face approach [TZS*16a] uses a compact set of only 76 expression coefficients, whereas for photo-real virtual humans in recent movie productions hundreds of blendshapes are employed for fine-scale control. Therefore, many recent approaches try to go beyond this coarse reconstruction to further personalize the model and recover the missing dimensions. In the following, we will give a detailed overview of these approaches.

7.1. Personalized Reflectance Maps

The low-dimensional statistical prior [BV99], which is employed by many state-of-the-art tracking approaches, models the skin reflectance variation of 200 3D scans (mostly Caucasians) using principal component analysis. Thus, it does not generalize well beyond the span of this ethnic group. If a colored SH illumination model is assumed, the unexplained skin color variation is typically baked into the ambient illumination coefficients during face reconstruction. Even though this is not physically correct, these cases are in general still robustly handled and do not lead to a degradation of the tracking quality. But this physically implausible explanation of skin color variation through illumination will cause errors in relighting and illumination transfer applications.

One of the biggest problems of this low-dimensional reflectance subspace is the lack of facial hair, e. g., beards. This leads to reconstruction and tracking artifacts, manifested as global geometric misalignments and surface shrinkage, since the model avoids to correctly overlay unexplained regions to reduce the fitting error.

A simple extension to the coarse albedo estimation based on a statistical prior is to compute a personalized reflectance map. This technique is used by many current approaches [GZW*16, IBP15, TZN*15, TZS*16a, TZS*16b]. First, the parameters of the coarse parametric model and the illumination are jointly estimated, thus recovering the facial identity and the incident illumination. The color information in the input image is projected to the model and divided by the estimated illumination. The resulting reflectance map contains high-frequency details, which significantly improves tracking accuracy. Note that this strategy only succeeds if the coarse alignment can be computed reliably.

7.2. Multi-region and Sparse Deformation Models

Researchers also looked at dividing the face into smaller regions in order to approximate faces better. Already Blanz and Vetter [BV99] partitioned the face into 5 regions, fitted a parametric model to each region independently and finally blend the estimates into a single coherent face shape. Since then, different methods for manual [DM00, TDlTM11] and automatic [JTDP03, DSVG10, NVW*13] face segmentation have been proposed. Most segmentation layouts attempt to split the face into few rather large and semantically meaningful regions at a single scale. Brunton et al. [BBW14] employ multi-linear wavelets to statistically analyze a face at multiple scales. Wu et al. [WBGB16] split the face into very many small patches and further increase the flexibility of the model by explicitly decoupling the rigid pose of the patch from its non-rigid

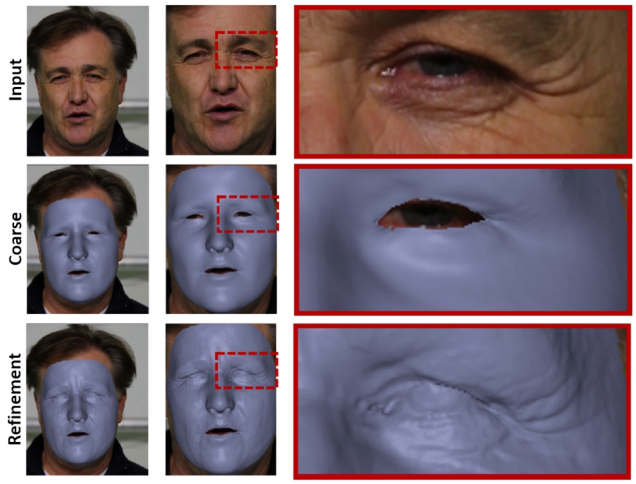

Figure 3: *The facial performance capture approach of Garrido et al. [GZC*16] employs shape-from-shading techniques to reconstruction wrinkle-level surface detail. Image taken from [GZC*16].*

deformation. In order for such a highly localized model to be applicable to monocular capture they employ anatomically inspired constraints that add the required regularization. It is also possible to automatically compute sparse localized deformation components from sequences of face meshes [NVW*13], either fully-automatically or under user guidance, instead of using a pre-defined global or local region model. Local models have the advantage of better generalization to unseen data, since they provide more degrees of freedom. On the down side, such models are generally more difficult to fit to noisy, monocular RGB and RGB-D input due to the limited spatial support of the underlying basis.

7.3. Medium-level Shape Correctives

Face reconstruction and tracking quality can be improved by integrating person-specific shape correctives on top of the generic face models. In the following, we focus on medium-level correctives, the recovery of small scale detail is discussed in the next section.

Li et al. [LYYB13] first proposed an incremental PCA refinement for on-the-fly personalization of the expression basis. They employ an RGB-D sensor and model the out-of-subspace motion based on Laplacian deformation. The final personalized expression basis leads to an improved face tracking quality. Bouaziz et al. [BWP13] proposed the use of a medium-scale corrective basis, which is based on manifold harmonics [VL08]. Manifold harmonics are a generalization of the Fourier basis to the mesh domain. They jointly fit this additional linear basis on top of the coarse-scale model based on RGB-D data. Recently, Garrido et al. [GZC*16] proposed an approach that recovers medium-scale correctives based on face video recorded with a single color camera.

7.4. Fine-Scale Detail Estimation

To be able to also recover small surface detail, shape-from-shading techniques can be applied for face reconstruction [Hor75, ZTCS99, WWMT11, WZN*14, GVWT13, SSS14, SWTC14] (see Fig. 3). These approaches exploit shading cues in the input color image to refine the surface geometry. The disadvantage of shading-based refinement is its high computational complexity, which makes it currently unsuitable for real-time techniques.

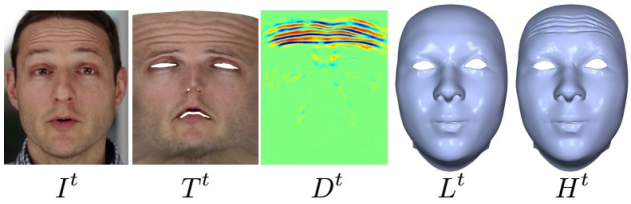

Figure 4: *[CBZB15] uses the RGB input I^t to compute a texture map T^t based on the coarse mesh L^t. A local regressor estimates the displacement map D^t to produce a high-fidelity mesh H^t. Image taken from [CBZB15].*

High-quality multi-view approaches obtain high-quality fine-scale surface reconstructions [BBA*07, FP09, BHPS10, BHB*11, FJA*14] based on a controlled capture setup at slow off-line frame rates. Based on a corpus of high-quality 3D scans, generative wrinkle formation models can be learned [HYZ*12, BBB*14, LXC*15a, CBZB15] as an alternative to shape-from-shading. High-resolution static face scans can be dynamically combined based on video input to model wrinkle formation [FJA*14]. Bickel et al. [BLB*08] propose pose space deformation by parameterizing a generative wrinkle model using a coarse feature graph. Such models even allow for the regression of wrinkle level detail at real-time frame rates [CBZB15] based on commodity data (see Fig. 4). Transient detail can also be learned using a data-driven approach [HYZ*12, MJC*08] to infer facial detail maps. The approach of Garrido et al. [GZC*16] learns a gradient domain regressor that models the correlation between transient surface detail and the low-dimensional blendshape expression coefficients.

7.5. Reconstruction of Personalized Face Rigs

Some state-of-the-art approaches go beyond the reconstruction of facial identity and coarse expression in monocular video. Their goal is rather the reconstruction of person-specific face rigs, i. e., fully-controllable and personalized parametric models that capture the mannerisms of a subject, such as personalized smiles and frown lines. Parameters of these rigs can be modified to plausibly reproduce even face expressions that were not seen in the videos from which they were reconstructed.

Face rigs can be created using either detailed tailor-made expressions [IBP15, LWP10, WLVGP09], physically-based geometric deformations driven by simulated muscle activations [EBDP96, SSRMF06], or a combination of both [KMML10]. These rigs are usually animated through high-level controllers, and they are employed in VFX pipelines and for various animation applications. In the literature, face rigs are commonly modeled as a combination of personalized blendshapes due to their simplicity and intuitive semantics. In this category, we can find methods that either adapt [LWP10, IBP15] or extend [LYYB13] the linear expression

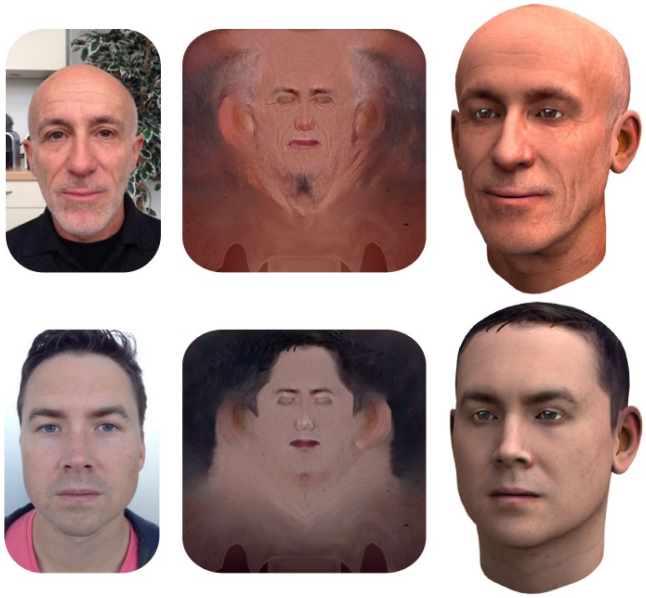

Figure 5: *[IBP15] create fully rigged 3D avatars from uncalibrated RGB data. The avatars are augmented with textures and dynamic detail maps that are used to generate wrinkles. Image taken from [IBP15].*

basis of a generic blendshape rig. Li et al. [LWP10] transfer example target expressions to a generic blendshape rig while preserving the original model semantics. Ichim et al. [IBP15] non-rigidly deform the generic blendshapes to optimally match a sequence of actor-specific expressions. In addition, they learn a linear mapping between fine-scale details and facial feature strain to add wrinkles to the synthesized expressions (see Fig. 5). Li et al. [LYYB13] extend the generic blendshape model by adding unobserved expressions to the model. Alternatively, Garrido et al. [GZC*16] learn a fully-controllable, personalized blendshape rig that models the correlation between the generic expression basis (represented as blendshape weights) and person-specific idiosyncrasies at a medium- and a fine-scale detail layer. Here, the fine-scale detail layer represents transient wrinkle-level detail. The reconstructed personalized face rigs allow artists to generate novel person-specific expressions with fine-scale wrinkles by using intuitive high-level controllers.

7.6. Anatomical Models and Physics

Finally, also non-linear face models can be applied [Law07] for face capture. In particular physical and anatomical models have attracted the attention of the research community in the recent years. Sifakis and colleagues [SNF05] created a full anatomical model of a human head from MRI data, which they used to fit to sparse MoCap markers, and recently Ichim et al. [IKKP17] explored anatomically based performance capture from a single RGB-D input. Creating a full anatomical rig can be very labor intense, and hence researchers have looked into ways to transfer and personalize a template rig [CBB*15]. Another line of research investigates anatomically inspired rigs. The work of Beeler and Bradley [BB14] employs empirically designed constraints that model the relationship between the skull and skin surface to estimate the exact pose of the skull given observations of the non-rigidly deforming skin - a process called stabilization. Wu et al. [WBGB16] explore the relationship of skin and skull further and propose a highly localized face model that is both flexible due to its locality and robust due to the global regularization provided by the skull.

8. Estimation of Model Parameters

There are two general approaches to estimate the parameters of a face model. An estimation approach is either generative or discriminative. Generative approaches fall back to the principle of analysis-by-synthesis where the model is adapted in an iterative manner until the synthesized face matches the input data. The analysis-by-synthesis technique is based on parameter optimization to minimize the difference between the synthetic and the observed image ($\Rightarrow$ Energy Minimization). As an alternative to optimization based approaches, parameter regression can be used (see Sec. 8.5). This report is focused on approaches in which optimization based reconstruction is the central concept. In the following, we describe the foundations of optimization based reconstruction in detail.

8.1. Parameter Estimation as Energy Minimization

Given a monocular video as input, parameter estimation is commonly phrased as a general non-linear optimization problem [LYYB13, BWP13, CWS*13, CWLZ13, SWTC14, SSS14, CHZ14, GZC*16, TZN*15, TZS*16a, WSXC16, TZS*16b]. To this end, first all unknowns are stacked in a parameter vector $\mathcal{P}$. Recent state-of-the-art approaches typically optimize for all or a subset of the following parameters: Global head rotation $\mathbf{R}$ and translation $\mathbf{t}$, facial shape $\{\alpha_i\}_{i=1}^{m_s}$, reflectance $\{\beta_i\}_{i=1}^{m_r}$, expression $\{\delta_i\}_{i=1}^{m_e}$ and illumination $\{l_i\}_{i=1}^{m_l}$.

The optimal parameters $\mathcal{P}^*$ are found by minimizing a reconstruction/tracking objective function E:

$$\mathcal{P}^* = \underset{\mathcal{P}}{\operatorname{argmin}} E(\mathcal{P}) \quad . \tag{7}$$

In general, the used tracking objectives are highly non-linear in the unknown parameters and consist of all or some of the following different components:

$$E(\mathcal{P}) = \underbrace{w_{dense}E_{dense}(\mathcal{P}) + w_{sparse}E_{sparse}(\mathcal{P})}_{data} + \underbrace{w_{reg}E_{reg}(\mathcal{P})}_{prior} \quad . \tag{8}$$

The first components are the data terms that measure the similarity of the current synthesized face and the input data. Several data terms are commonly employed: Sparse feature alignment terms E_{sparse} match a small number of feature points of the model to equivalent detected feature points in the input. Additionally, dense alignment terms are often employed that enforce constraints on vertex [LYYB13, BWP13, GVWT13, SWTC14, SSS14, GZC*16] or on pixel level [CWS*13, TZN*15, TZS*16a, TZS*16b]. In addition to the data terms, regularizers E_{reg} are employed to better constrain the solution and resolve ambiguities. These components are linearly combined together based on a set of weights $w_\bullet$. In the following, we discuss several common choices for the different components of the objective function. An overview of the influence of different alignment terms is shown in Fig. 6.

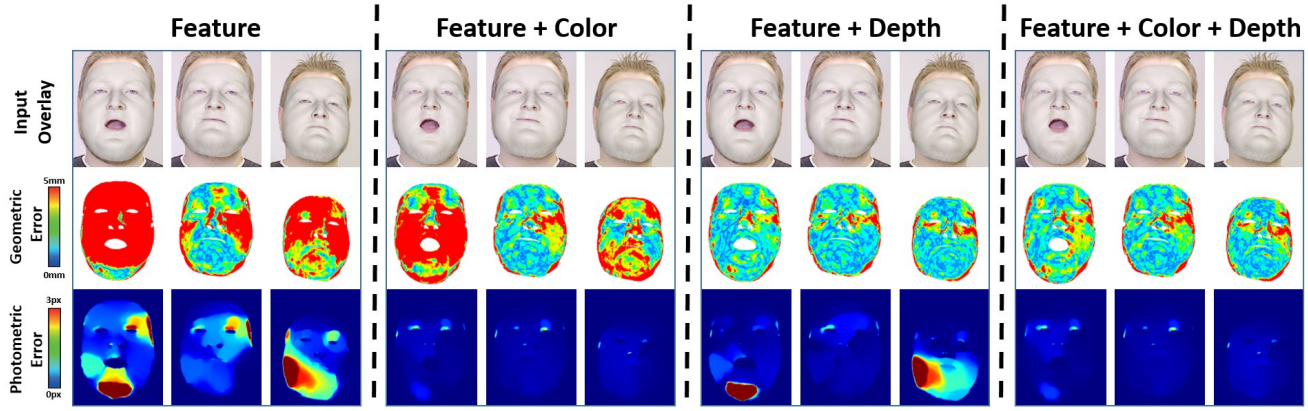

Figure 6: *Influence of the energy terms as shown in [TZN*15]: tracking accuracy is evaluated in terms of geometric (middle) and photometric error (bottom). The final reconstructed pose is shown as an overlay on top of the input images (top). Mean and standard deviations of geometric and photometric error are 6.48mm/4.00mm and 0.73px/0.23px for Feature, 3.26mm/1.16mm and 0.12px/0.03px for Features+Color, 2.08mm/0.16mm and 0.33px/0.19px for Feature+Depth, 2.26mm/0.27mm and 0.13px/0.03px for Feature+Color+Depth. Image taken from [TZN*15].*

8.1.1. Sparse Feature Alignment

Since faces contain many visually-salient structures, one of the most commonly used and most important data terms is sparse feature alignment [VBPP05, WBLP11, CWS*13, LYYB13, HMYL15, CBZB15, SLL16, CWW*16, SWTC14, SSS14, GZC*16, TZN*15, TZS*16a, TZS*16b]:

$$E_{\text{lan}}(\mathcal{P}) = \frac{1}{|\mathcal{F}|} \sum_{\mathbf{f}_j \in \mathcal{F}} w_{\text{conf},j} \left\| \mathbf{f}_j - \mathbf{K}\Pi(\mathbf{R} \cdot \mathbf{v}_j(\mathcal{P}) + \mathbf{t}) \right\|_2^2 . \quad (9)$$

It enforces that a small set of facial landmark points $\mathbf{v}_j(\mathcal{P})$ on the model align well with corresponding detected 2D features $\mathbf{f}_j \in \mathcal{F} \subset \mathbb{R}^2$. Each of the detections normally comes with a confidence $w_{\text{conf},j}$ that can be used to weight the corresponding alignment constraint.

There is a rich body of literature on methods for the detection and tracking of facial features and landmarks. Some methods are holistic, using the whole face region to localize the facial landmarks like Active Appearance Models (AAM) [CTCG95, CET01, MB04, XBMK04]. Local methods often use constraints to stabilize the estimation [SLC11a, CC06, BRM12, AZCP13]. These constrained localized model (CLM) approaches are fast and robust. Therefore, they are commonly used as facial feature tracker for dense face tracking and reconstruction. For a good overview of face detection in the wild see [VJ04, ZZZ15].

Given the fact that the energy minimization problem is highly non-linear, E_{lan} is a very important energy term. The facial feature points are not only important for initializing the tracker in the first frame, but also bring the optimization close to the basin of convergence of the dense alignment constraints. Without this data term fast motion and strong deformation can not be tracked reliably.

8.1.2. Dense Photometric Alignment

The sparse data term coarsely aligns the model to the input, such that the optimization is in the basin of convergence of the dense alignment constraints. Dense photometric alignment is commonly used by many approaches [GVWT13, SSS14, TZN*15, CBZB15, TZS*16a, TZS*16b, GZC*16, WBGB16]. The photometric alignment term measures densely how well the input data is explained by a rendered version of the current fit:

$$E_{\text{col}}(\mathcal{P}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} \| C_{\mathcal{S}}(\mathcal{P}, \mathbf{p}) - C_{\mathcal{I}}(\mathbf{p}) \|_2^2 , \quad (10)$$

In this formulation, $C_{\mathcal{S}}$ is the synthesized face image, $C_{\mathcal{I}}$ is the input RGB image, and $\mathbf{p} \in \mathcal{V}$ denotes all visible pixel positions in $C_{\mathcal{S}}$ that are used for comparison. For performance reasons this constraint is often defined on vertex level rather than pixel or fragment level. Different norms are used to measure closeness. Normally, either an ℓ_2 [BV99, GVWT13, SSS14, TZN*15, GZC*16] or a more robust $\ell_{2,1}$-norm [TZS*16a, TZS*16b] is employed. The latter norm handles outliers better. In the $\ell_{2,1}$-norm, the distance in color space is based on ℓ_2, while the summation over all pixels enforces sparsity based on an ℓ_1-norm.

8.1.3. Dense Geometric Alignment

Besides the dense photometric constraints, all RGB-D based approaches rely on a dense geometric alignment term that exploits the available depth stream to better constrain the reconstruction problem and resolve depth ambiguity. This is based on the idea that the reconstructed geometry of the face model should tightly align to the captured depth stream [WBLP11, CWS*13, LYYB13, BWP13, TZN*15, HMYL15, TZS*16b]. To this end, multiple surface-to-depth distance measures can be employed. The most commonly used alignment term measures the point-to-point distance between model surface points and the input depth. This constraint is effectively the sum of the projective Euclidean point-to-point distances to the input:

$$E_{\text{point}}(\mathcal{P}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} \| X_{\mathcal{S}}(\mathcal{P}, \mathbf{p}) - X_{\mathcal{I}}(\mathbf{p}) \|_2^2 , \quad (11)$$

with $X_{\mathcal{S}}(\mathcal{P},\mathbf{p})-X_{\mathcal{I}}(\mathbf{p})$ being the difference between the measured 3D position and the corresponding 3D model. $\mathcal{V}$ denotes all visible pixel positions that are used for comparison. Often times, this constraint is defined on vertex level rather than a per (depth-)pixel level.

To improve robustness and convergence a first-order surface approximation distance [CM92] is often used. This point-to-plane surface metric is especially beneficial for translational motion, where a simple point-to-point metric would fail:

$$E_{\text{plane}}(\mathcal{P}) = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{p}\in\mathcal{V}}\left[N_{\mathcal{S}}(\mathcal{P},\mathbf{p})^{\top}\cdot(X_{\mathcal{S}}(\mathcal{P},\mathbf{p})-X_{\mathcal{I}}(\mathbf{p}))\right]^2 , \quad (12)$$

with $N_{\mathcal{S}}$ being the synthesized surface normal image. One question that arises is if one should use a point-to-plane distance from input to model or vice versa. The advantage of using the normals of the model is that they are guaranteed to be smooth. Some approaches [TZN*15] propose to use a symmetric point-to-plane distance that goes from input to model and vice versa.

8.1.4. Regularization

Using the statistical information available in the face prior, the reconstruction problem can be further constrained to resolve ambiguities:

$$E_{\text{reg}}(\mathcal{P}) = \sum_{i=1}^{m_s}\left(\frac{\alpha_i}{\sigma_{s,i}}\right)^2 + \sum_{i=1}^{m_r}\left(\frac{\beta_i}{\sigma_{r,i}}\right)^2 + \sum_{i=1}^{m_e}\left(\frac{\delta_i}{\sigma_{e,i}}\right)^2 . \quad (13)$$

This constraint enforces the model parameters to be statistically close to their mean. This is a commonly-used regularization strategy [BV99, TZN*15, TZS*16a, TZS*16b, GZC*16] that prevents degeneration of the facial geometry and face reflectance. Regularization is particularly important in the monocular RGB reconstruction setting, which suffers from severe depth ambiguity.

8.2. Model Initialization

While joint estimation of all parameters is in general feasible, almost all approaches fix the facial identity at a certain stage in the algorithm, and subsequently optimize only for the remaining subset of parameters. Many approaches use multiple images [WBLP11, LYYB13, IBP15, TZN*15, TZS*16a, GZC*16] at the beginning of an input sequence. This enables a better identity estimate than reconstruction from a single frame. A few examples of the different employed strategies are described in the following. Weise et al. [WBLP11] fuse multiple depth frames based on rigid alignment to obtain a 3D model of the neutral face. Thies et al. [TZN*15, TZS*16a] use model-based non-rigid bundle adjustment over 3 frames with different head pose. Ichim et al. [IBP15] perform a static structure-from-motion reconstruction based on multi-view data captured by an iPhone. The approach of [GZC*16] averages the identity over a small part of the video.

8.3. Handling Occlusions

Occlusion handling is of paramount importance for robust face tracking. Facial hair, a hand or other occluders can easily deteriorate the previously described face reconstruction and tracking approaches [LYYB13, BWP13, CWS*13, CWLZ13, SWTC14, SSS14, CHZ14, GZC*16, TZN*15, TZS*16a, WSXC16, TZS*16b], since they assume the absence of occlusion. Recently, many approaches have been proposed to alleviate this problem. In general, they are based on a segmentation mask that disables the data fitting terms in occluded regions. Hsieh et al. [HMYL15] use depth and color information of an RGB-D camera to robustly segment the visible face region (see Fig. 7). After segmentation, they apply occlusion completion based on the appearance of the last unoccluded observation of the face. Saito et al. [SLL16] use deep learning to obtain a clean facial segmentation of the visible part of the face based on just monocular RGB data (see Fig. 8). This binary mask is later used in their adapted displaced dynamic expression (DDE) method [CHZ14]. Such a skin mask could also be incorporated into dense reconstruction methods [TZN*15, TZS*16a] to weight the data terms, but it also requires an adaptation of the used sparse feature tracker, which is also adversely impacted by occlusions.

Figure 7: *[HMYL15] demonstrates a face tracking approach that is robust to occlusion. They use the depth and color information of an RGB-D camera to robustly segment the face region. Image taken from [HMYL15].*

Figure 8: *An RGB face tracker that is robust to occlusion has been proposed by [SLL16]. Using deep learning a clean segmentation of all visible parts of the face is obtained. Image taken from [SLL16].*

8.4. Optimization Strategies

One of the main components of an analysis-by-synthesis approach is the estimation of face parameters that minimize the difference between the input and the synthesized face image. In general, to find this parameter estimate an optimization problem has to be solved. An analysis-by-synthesis approach iteratively computes a new set of parameters $\mathcal{P}^{i+1}$, starting from the synthesis generated with the old parameters $\mathcal{P}^{i}$. The parameters $\mathcal{P}^{i+1}$ are chosen such that the energy that measures the difference between the observation and the synthetic image is minimized. For simplicity, let us assume an objective function $E(\mathcal{P})$ that is comprised of a single data fitting

term:

$$E(\mathcal{P}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} \Psi(\mathbf{r}(\mathcal{P}, \mathbf{p})) \ . \tag{14}$$

Here, $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ is a function that maps the difference $\mathbf{r}(\mathcal{P}, \mathbf{p}) = C_{\mathcal{S}}(\mathcal{P}, \mathbf{p}) - C_{\mathcal{I}}(\mathbf{p})$ between the observations and the model to a scalar. The number of residuals is often proportional to the number of pixels or the number of vertices. Note, the number of residuals is not necessarily a constant, i. e., it might change during the analysis-by-synthesis iterations.

The most common metric is the ℓ_2-norm ($\Psi(\mathbf{x}) = ||\mathbf{x}||_2^2$). In this case, the optimization problem boils down to a non-linear least-squares problem. Thus, Eqn. 14 can be rewritten to:

$$E(\mathcal{P}) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} ||\mathbf{r}(\mathcal{P}, \mathbf{p})||_2^2 = ||F(\mathcal{P})||_2^2 \ . \tag{15}$$

If $F(\mathcal{P})$ is a linear function in the model parameters $\mathcal{P}$, the optimization problem collapses to a linear least-squares problem ($||\mathbf{A}\mathcal{P} + \mathbf{b}||^2 \rightarrow \min$) that can be solved using the corresponding normal equations $\mathbf{A}^\top \mathbf{A}\mathcal{P} = -\mathbf{A}^\top \mathbf{b}$.

For most current state-of-the-art approaches, $F(\mathcal{P})$ is a non-linear function due to the assumption of a perspective camera. To solve the resulting non-linear least-squares problem, iterative methods such as gradient descent, *Gauss-Newton* or *Levenberg-Marquardt* are applied. Quasi-Newton methods, such as LM, often perform some kind of step size control, which requires the evaluation of the objective function after each iteration step. The *Gauss-Newton* algorithm is an approximation of *Newton*'s method. One advantage over *Newton*'s method is that no second order derivatives, which are often challenging to compute, are required. For an in detail discussion of such approaches we refer to [NW06].

Newton's method can be applied to general non-linear problems and iteratively calculates an updated solution:

$$\mathcal{P}^{i+1} = \mathcal{P}^i - \mathbf{H}_E(\mathcal{P}^i)^{-1} \cdot \nabla E(\mathcal{P}^i) \ .$$

Here, $\nabla E(\mathcal{P}^i)$ denotes the gradient and $\mathbf{H}_E(\mathcal{P})$ the Hessian of the energy function, thus, involving second order derivatives of the residual vector $F(\mathcal{P})$. The *Gauss-Newton* method is limited to non-linear least-squares problems and approximates the Hessian using only first order derivatives:

$$\mathbf{H}_E(\mathcal{P}) \approx 2 \cdot \mathbf{J}^\top(\mathcal{P}) \cdot \mathbf{J}(\mathcal{P}) \ .$$

Here, $\mathbf{J}(\mathcal{P})$ is the Jacobian of the residual function $F(\mathcal{P})$. Using this approximation, results in the following update rule:

$$\mathcal{P}^{i+1} = \mathcal{P}^i - \underbrace{\left[\mathbf{J}^\top(\mathcal{P}^i) \cdot \mathbf{J}(\mathcal{P}^i)\right]^{-1} \cdot \mathbf{J}^\top(\mathcal{P}^i) \cdot F(\mathcal{P}^i)}_{\Delta^i} \ . \tag{16}$$

Thus, to compute the parameter update Δ^i the following system of linear equations has to be solved:

$$\left[\mathbf{J}^\top(\mathcal{P}^i) \cdot \mathbf{J}(\mathcal{P}^i)\right] \cdot \Delta^i = \mathbf{J}^\top(\mathcal{P}^i) \cdot F(\mathcal{P}^i) \ . \tag{17}$$

These kinds of linear equations can be solved by iterative solvers (e.g., preconditioned conjugate gradient (PCG) as in [TZN*15, TZS*16a]). Considering the limited time budget of real-time approaches, iterative solution strategies have the advantage that they can trade accuracy vs. speed by reducing the number of iterations.

The *Gauss-Newton* method is an iterative algorithm which is highly dependent on the initial guess. Sparse detected landmarks can be used to initialize face tracking. This solution is then used as starting point for the consecutive frame.

To enable tracking of fast motion, hierarchical optimization strategies are employed. I.e., the observations are down-sampled (e.g., half the image resolution) and the optimization is first solved on such a coarse level before being propagated to the next finer level. This helps convergence, since a pixel in the down-sampled image covers a larger part of the scene. Thus, if an object moves for example four pixels, it moves only two pixels in the once and only one pixel in the twice down-sampled image. By optimizing first on a down-sampled image, the numerical derivatives of the observed image have a more global footprint and the residual function is smoother. This reduces the number of local minima and leads to faster convergence.

The *Levenberg-Marquardt* method [Mar63, Lev44, Mor78] is a mixture between *Gauss-Newton* and *Preconditioned Gradient-Descent*. It prevents the *Gauss-Newton* method from overshooting.

$$\mathcal{P}^{i+1} = \mathcal{P}^i - (\mathbf{J}^\top(\mathcal{P}^i) \cdot \mathbf{J}(\mathcal{P}^i) + \lambda \cdot \mathbf{D})^{-1} \cdot \mathbf{J}^\top(\mathcal{P}^i) \cdot F(\mathcal{P}^i) \ . \tag{18}$$

Here, $\mathbf{D} = \text{diag}(\mathbf{J}^\top(\mathcal{P}^i) \cdot \mathbf{J}(\mathcal{P}^i))$ is a diagonal matrix and λ is the factor that controls the weighting between *Gauss-Newton* and *Preconditioned Gradient-Descent* in addition to the step length. The factor λ is reduced if the update step reduces the energy function. If the update step would increase the energy, λ is increased until the energy function is reduced. Note, by increasing λ the algorithm favors the update direction of *Preconditioned Gradient-Decent* and reduces the step size.

Another common metric is the $\ell_{2,1}$-norm ($\Psi(\mathbf{x}) = ||\mathbf{x}||_2^1$). Using the Iteratively Reweighted Least Squares (IRLS) method, for more details see [Bjö96], the optimization problem can be transformed into a sequence of non-linear least-squares problems. The key idea of IRLS is to split the norm in two components:

$$||\mathbf{r}(\mathcal{P}, \mathbf{p})||_2^1 = \underbrace{(||\mathbf{r}(\mathcal{P}, \mathbf{p})||_2^1)^{-1}}_{constant} \cdot \underbrace{||\mathbf{r}(\mathcal{P}, \mathbf{p})||_2^2}_{least-squares} \tag{19}$$

The first part is assumed to be constant in each iteration step and the second part is in least-squares form, thus, can be optimized using a *Gauss-Newton* solver as described above.

8.5. Machine Learning for Dense Face Reconstruction

Besides the regression of a small number of control points [CWLZ13, CHZ14], which can be used to fit a parametric face model, machine learning based approaches are able to perform dense face reconstruction. Very recently, first approaches [MKB*16, OLSL16, KMB*17, LKA*17] have been proposed that try to bypass the costly optimization step by learning a regression function that directly estimates the model parameters from a given video sequence. Other approaches focus on monocular face reconstruction from a single input image [RSK16, RSOK17, TZK*17,

SRK17, TTHMM17]. Most of these approaches employ either Random Forests [MKB*16, KMB*17] or Deep Convolutional Neural Networks (CNNs) [OLSL16, LKA*17, RSK16, RSOK17, TZK*17, SRK17, TTHMM17] to learn the image-to-parameter or image-to-geometry mapping directly from a training corpus. While this report is focused on optimization-based approaches, the results obtained by these learning-based techniques are inspiring, and we expect to see heavy use of techniques based on deep learning in the future. The training corpus is normally based on reconstructed ground truth face geometry [TTHMM17, LKA*17], fully synthetically generated images [RSK16, RSOK17, SRK17], a mixture of the previous two approaches [MKB*16, KMB*17], or just consists of in-the-wild images without available ground truth [TZK*17]. Richardson et al. [RSK16] use shading-based surface refinement on top of a coarse regression result to recover fine-scale surface detail. In a follow-up work [RSOK17], surface refinement is also phrased as a learning task. The network is trained end-to-end for regressing coarse shape and a fine-scale detail layer. The approach of Tewari et al. [TZK*17] tightly integrates deep learning-based and model-based capture in an end-to-end trainable architecture based on an expert-designed differential image-formation layer that enables unsupervised training on in-the-wild face images. In contrast to the previous approaches, this approach jointly estimates dense surface geometry, reflectance, and scene illumination. Face shape and texture estimation can be used for face identification [TTHMM17]. The approach of Sela et al. [SRK17] is trained fully synthetically and regresses per-pixel depth instead of model parameters. This enables the approach to leave the restricted linear subspace spanned by the underlying model used for data generation, leading to more detailed reconstruction results. While regression based approaches work well in practice and deliver impressive results at high-speed, optimization-based techniques are still able to obtain slightly better reconstruction quality and are not biased by the used training corpus, i.e., they generalize better to difficult situations. In the future, we expect hybrid approaches that combine machine learning techniques for parameter initialization with their optimization-based counterparts to exploit the best of both worlds; thus the optimization-based approach is initialized close to the optimum, which guarantees fast convergence.

9. Beyond Face Reconstruction

In previous sections, only a tightly confined facial mask region has been considered [WBLP11, LYYB13, CWS*13, BWP13, SWTC14, SSS14, HMYL15, TZN*15, SLL16, TZS*16a, GZC*16], but there has been extensive work on extending reconstruction and tracking approaches to go beyond the face mask. In the following, we present a detailed discussion of approaches that focus on the reconstruction of individual parts of human heads.

9.1. Lip Reconstruction

An accurate and robust reconstruction of high-quality lip shapes is of paramount importance for realistic and believable virtual characters, e.g. for movies and computer games, since even small differences in mouth shape drastically influence the interpretation of speech and the conveyed emotion. One famous example that shows the tight coupling between the visual and audio channel, when it comes to the perception of speech, is the so-called McGurk effect [MM76]. This effect demonstrates that a video with a modified lip motion, but the same audio track, can make a viewer perceive a different vowel compared to the unmodified video clip.

Human lips are challenging to reconstruct and track, since they exhibit an incredible range of motion and deformation, such as in a kiss. Furthermore, the inner region of the lips is typically hidden by occlusion, and the lips exhibit drastic appearance changes due to blood flow and wetness. These challenges make lip motion hard to capture with current state-of-the-art monocular face tracking approaches [WBLP11, LYYB13, CWS*13, BWP13, SWTC14, TZN*15, SSS14, TZS*16a, GZC*16, WSXC16], especially if the motion contains significant amounts of stretching, bending and rolling. Even approaches that use multi-view input fail to track the lip motion correctly leading to implausible reconstruction results. First approaches [BHPS10, BGY*13] that address this challenging reconstruction problem require dense multi-view capture under controlled illumination and integrate additional contour cues into the optimization. Other approaches [ASC13] use multi-camera photometric stereo to match lip contours with predefined iso-lines on the model. The iso-lines are used to tackle the problem of sliding contours, e.g., for the inner contour line of the lips, since there exists no constant point-to-point association between the detected 2D contour and the model.

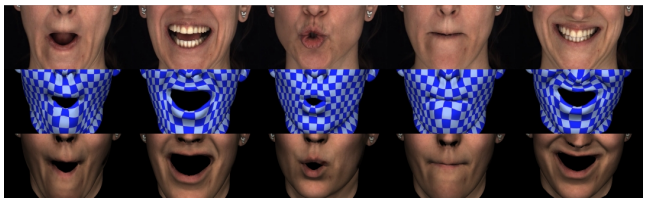

Figure 9: *A monocular lip reconstruction method recovers complex lip shapes like stretching, bending and rolling of the lips using a learning-based approach [GZW*16]. Image taken from [GZW*16].*

Recently, a first learning-based monocular reconstruction approach [GZW*16] has been proposed that successfully deals with stretching, bending, and rolling of the lips (see Fig. 9). This approach employs a robust gradient domain regressor to estimate lip shape correctives with respect to a coarse state-of-the-art dense face tracking approach [GZC*16]. The regressor is trained to predict the difference between inaccurate lip shapes estimated by the coarse-scale face tracker and the high-quality multi-view reconstructions of Beeler et al. [BHB*11] based on a training corpus.

9.2. Mouth Interior, Teeth and Tongue Capture

In the graphics literature there is not a lot of work on capturing and modeling the mouth interior based on visual input. Similar to lips, the mouth interior is hard to reconstruct and track from external visual data due to severe occlusions and complex material properties, e.g., teeth exhibit subsurface scattering and are highly glossy. Recent monocular reconstruction approaches rely on generic models of the mouth interior [CWW*16, IBP15], which are statically rigged to the underlying blendshape basis.

Other approaches employ either an image-based mouth synthesis approach [TZS*16a] or only render a coarse geometric textured proxy [GVS*15, TZN*15]. Recently, an approach was proposed that can capture high-quality teeth geometry from external multi-view images, using a statistical prior of teeth structure built from dental scans [WBG*16] (see Fig. 10). In related research fields, modeling the tongue based on MRI data [HSW14, HSB*16] has been examined. To date, however, no method has attempted to combine facial performance capture with high-quality reconstruction and tracking of all inner mouth components in a unified system.

Figure 10: *Despite the lack of visual features on the teeth, Wu et al. [WBG*16] present a statistical model of teeth shape and a method to fit it to multi-view imagery. Image taken from [WBG*16].*

9.3. Capturing Eyes and Eyelids

An exploration of facial capture algorithms would not be complete without discussing eyes. Arguably, eyes can be considered the most important feature of a face, as this is where we focus most of our attention when looking at another person. Traditional face capture methods tend to ignore eyes, due to their complex appearance properties which violate the typical assumptions that many reconstruction algorithms rely on. For example, eyes are very specular, they are often largely occluded by the surrounding skin, and the eyeball itself is composed of several components including the translucent white sclera, the fully transparent cornea in front, and a volumetric colorful iris that contracts and expands to control the pupil size. For these reasons, the eye region is typically treated quite separately when it comes to reconstruction, tracking and animation.

Eye tracking is actually a fairly old topic, as early methods were interested in determining the point on a 2D display where a user is looking, often using specialized hardware (e.g. infrared light). Hansen and Ji [HJ10] provide an extensive survey of the first eye models and algorithms used for tracking gaze in images. From the perspective of 3D facial capture, 2D gaze tracking is generally not sufficient. More recent research has focused on capturing and tracking 3D eyes, with a focus on virtual character modeling and animation. A first approach for capturing the detailed 3D structure of eyes was presented by Bérard et al. [BBN*14] (see Fig. 11 (a)). Their method uses a combination of algorithms to recover all the visible parts of an eye, including the sclera, cornea and iris. The main drawback of their method is that the capture setup is quite involved, and the capture process is cumbersome and uncomfortable for the actor. To alleviate this issue, Bérard et al. followed up with a more lightweight capture solution, able to achieve similar eye reconstructions from a single photo as input [BBGB16] (see Fig. 11 (b)). This lightweight approach utilizes a database of high-quality 3D eyes captured from their original multi-view system, and builds a parametric eye model that can then be fitted to new images of unseen eyes. Another example of model fitting for eye reconstruction was presented by Wood et al. [WBM*16], who create a parametric model of the eyeball coupled with a 3D morphable model of the facial eye region, then proceed to fit the models to images using analysis-by-synthesis. In the realm of online facial tracking, Wang et al. [WSXC16] have proposed the first method for combining real-time facial reconstruction with 3D eye gaze capture from video input (see Fig. 12).

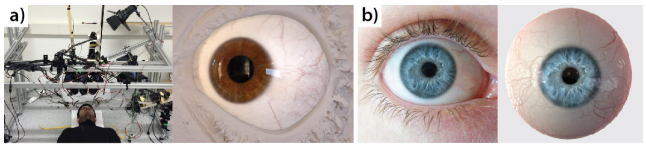

Figure 11: *Bérard et al. have studied algorithms for high-quality capture of eyes, first using a detailed capture setup (a) [BBN*14], and then with a lightweight approach that can operate on a single image (b) [BBGB16]. Images taken from [BBN*14] and [BBGB16].*

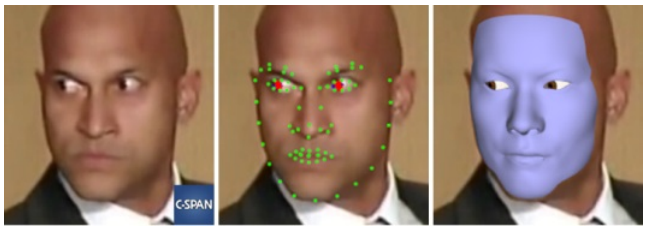

Figure 12: *Wang et al. [WSXC16] show real-time facial reconstruction in combination with 3D eye gaze capture from RGB video. Image taken from [WSXC16].*

Nearly as difficult to capture as the eyes themselves is the skin region around the eyes. Often partially occluded by eyelashes or completely invisible in an eyelid fold, the subtle nuances of this face region often go undetected by traditional face capture methods. The main exception is the work of Bermano et al. [BBK*15], who present a spatio-temporal reconstruction algorithm specifically tailored to recover the unique deformation behavior of eyelids. Coupled with existing face reconstruction methods, this technique provides a more complete facial tracking solution.

Recently, a technique [WXLY17] for tracking the 3D shape and motion of eyelids in real time has been proposed. The results are tightly integrated with the tracked face and eyeball. This approach represents eyelid variation using two linear spaces and performs semantic edge detection using a DNN, and handles partial detections using polynomial fitting.

9.4. Hair Capture

While hair is typically not considered part of the face, a person's hairstyle can be an important identifiable feature, nearly as

unique as the face itself. Among the first high-quality hair capture algorithms was the seminal work of Paris et al. [PCK*08] who designed the "hair photobooth", a method for capturing the shape and appearance of a hairstyle using an active lighting approach with multiple projectors and cameras. This innovation led to several follow-up methods for capturing static hairstyles [HZW12, HMLL14, ZCW*17], even those with complex braided patterns [HML*14], and stylized hairstyle capture for generating 3D printed figurines [EBGB14].

More challenging than static hair capture is to reconstruct dynamic hair in motion. This problem is particularly difficult due to the complex interaction of hair fibers, including self-contact, friction, adhesion, and occlusion. Nevertheless, the problem has been studied and initial results are inspiring. Luo et al. [LLP*12] present a multi-view system that relies primarily on orientation fields to recover hair structure, and show that the approach can be used on dynamic hair, although without establishing correspondence over time. The state-of-the-art in capturing temporally-coherent dynamic hair is the method of Xu et al. [XWW*14], who formulate the problem as a spacetime optimization over the projected motion path of local hair strands. The reconstructed geometry is impressive, however the hair animation can only be played as captured without an easy way to edit or re-direct the animation. To allow for artist edits post-capture, Derouet-Jourdan et al. [DJBDDT13] turn to physical simulation and aim to capture the physical properties of a hairstyle that can be animated in a directable way through simulation. Hu et al. [HBLB17] extend this idea for dynamic sequences of hair, presenting a framework for simulation-ready hair capture that can be applied to full hairstyles in motion, independent of the particular simulation method.

Hair modeling from a single view has also gained popularity. Chai et al. [CLS*15] recover depth maps from portrait photos using a morphable model for the face and shape from shading for the hair structure, regularized by a 3D helical hair prior. Hu et al. [HMLL15] use a 3D hairstyle database and reconstruct hair from a single image guided by a few user strokes. The latest work by Chai et al. [CSW*16] demonstrates fully automatic hair reconstruction from a single image, based on a deep learning approach.

In the context of facial capture, perhaps most relevant is the method of Beeler et al. [BBN*12] who couple high-quality face reconstruction with facial hair scanning into a single system. What is yet to be investigated is dynamic reconstruction of facial hair coupled with facial performance capture. Current literature also lacks high-quality real-time capture techniques for hairstyles, or monocular dynamic capture methods for hair. Although real-time methods for simulating hair have been presented [CZZ14], so real-time hair capture guided by physical simulation may not be far off. A selection of algorithms for capturing hair are illustrated in Fig. 13.

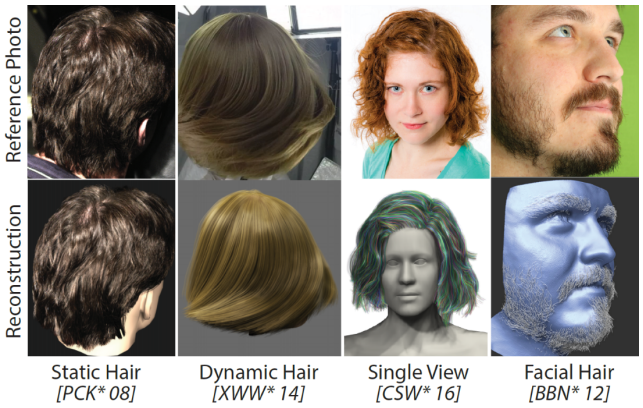

Figure 13: *Algorithms for capturing hair have focused on static hairstyles [PCK*08], dynamic hairstyles [XWW*14], single-view hair reconstruction [CSW*16], and facial hair capture [BBN*12]. Images taken from [PCK*08], [XWW*14], [CSW*16], and [BBN*12].*

9.5. Reconstruction of Complete Heads

The ultimate goal of face reconstruction is to capture a full high-quality personalized model of the entire human head, ready for use in movie and game production, from just a single monocular input image. Recent research has delivered impressive approaches for high-quality capture and tracking of face mask regions, and individual parts of the face, such as lips, teeth, eyes, eyelids and hair. Most of these approaches still require multi-view input and controlled setups, and are specifically tailored to only solve a small portion of the complex joint reconstruction problem. Currently, bringing all these individual components together to obtain a ready-to-use and complete head model, often desired as a parametric face rig, is still a time consuming and tedious process that involves tremendous manual work by artists. Especially, if photo-real virtual humans are the requirement [ARL*10, AFB*13, SEL17], a fully automatic solution has not been achieved so far.

Recently, the FLAME shape and expression model [LBB*17] has been released which has been constructed from thousands of 4D scans. It is the first available model that models the head and neck together, employs pose-dependent blendshapes that capture how the neck deforms during rotation, has an articulated jaw, and an eye model.

Several methods have made progress in reconstructing complete virtual heads. As mentioned earlier, the stylized hair capture method of Ecchevarria et al. [EBGB14] enables reconstruction of a complete head including the hairstyle for fabrication purposes. Based on multi-view data captured by a mobile phone, Ichim et al. [IBP15] reconstruct a full head model and afterwards personalize the expression dimensions based on a calibration sequence. Liang et al. [LSKS16] reconstruct the entire head geometry from a photo collection. Cao et al. [CWW*16] jointly reconstruct image-based dynamic avatars that contain parts of the upper body and hair based on coarse geometric proxies and a calibration sequence. Afterwards, tracking can be performed from monocular video alone and runs at real-time frame rates.

Recently, Hu et al. [HSW*17] proposed an approach that estimates a complete 3D avatar, including hair and teeth, from just a single image. They do not aim for photo-real reconstruction but for stylized capture of the head that can for example be used in games.

All of these approaches, while taking a large step forward, are still far from delivering a complete, anatomically correct and physi-

cally accurate head model that is ready for use in production. Bridging this content creation gap will be of paramount importance, especially for many virtual and augmented reality applications.

10. Applications of Face Reconstruction

Dense face reconstruction and tracking based on lightweight capture setups are an enabling technology for many exciting applications. In the following, we review several of those applications and discuss their requirements with respect to the underlying facial performance capture approach.

10.1. Facial Puppetry

Facial puppetry or cloning is a form of video-driven facial animation that is employed in the movie and game industry as well as teleconferencing to animate virtual characters. Here, the goal is to transfer expressions and emotions of a user in an input video stream to a target character rig.

Expressions can be cloned via semantically meaningful animation parameters or by mapping dense motion fields between the user and the target rig. Parameter-based methods directly transfer parameters between source and target. These methods either assume that both the source and target model share the same expression basis [VBPP05, WBLP11, BWP13, CWLZ13, LYYB13, CHZ14] or learn a linear subspace between the models to compensate for global shape differences during transfer [SLC11b, TMM*09, WLVGP09, GVS*15, TZS*16a]. Motion-based methods map the dense 2D facial motion of the source video onto the 3D geometry of the target rig [CXH03, NN01, SDT*07, SSK15] provided that dense correspondences are available. The approach of [RZL*17] automatically retargets the facial animation from an actor to a stylized character. The transfer is based on an automatically computed mapping that exploits the similarity between the motion space spanned by the blendshapes and an expressive calibration sequence of the actor. This motion mapping is more natural than classical deformation transfer.

In the literature, we can also distinguish between off-line and online approaches. Online methods target teleconferencing scenarios and focus on driving target rigs (usually non-human avatars) with coarse-scale facial deformation or motion parameters to produce plausible animations [BWP13, CWLZ13, CHZ14, CXH03, LYYB13, SLC11a, WBLP11] (see Fig. 14). Thus, they offer a good trade-off between accuracy and real-time performance. In contrast, off-line methods drive photo-realistic target rigs that preserve the face shape, details and mannerisms of the target character [IBP15, GZC*16, SSK15]. These methods sacrifice real-time performance to obtain higher accuracy and realism. However, data-driven approaches have demonstrated that real-time facial puppetry at high-fidelity can also be achieved by either learning a linear mapping between user expressions and high-quality target rig deformations [WLVGP09] or by constructing a target database of detailed subject-specific deformations that correlate to generic facial expressions performed by the user [CWW*16, BBB*14]. Video-driven facial animation of a subject's scanned 3D head model has also been applied to model-based coding [EG98] for video compression in a teleconferencing setting.

Figure 14: *Cao et al. [CHZ14] and Bouaziz et al. [BWP13] demonstrate facial puppetry systems that drive digital avatars in real time. Images taken from [CHZ14] and [BWP13].*

10.2. Face Replacement

Several approaches for face replacement or swapping have recently been proposed in the literature. Such methods replace the face of an actor in a target video with the face of a source actor, where the videos may be recorded under different illumination conditions. The main challenge is to synthesize a novel sequence that looks realistic and maintains temporal consistency. Here, the synthesized sequence preserves either the source or target performance. Fig. 15 shows an example of face replacement.

Figure 15: *The face replacement systems of [DSJ*11], [GVR*14] and [TZS*16a]. Note that Face2Face [TZS*16a] keeps the target face identity and adapts the expression according to the source. Image from [TZS*16a] (supplemental material).*

In the literature, we can find either model-based [JCG*08, DSJ*11] or pure image-based [GVR*14] approaches to tackle this challenging problem. Model-based methods first track the 3D pose and facial attributes (e. g., identity, expression and alternatively intrinsic face properties) of both the source and target face. The source face is then rendered under target conditions. A similar strategy was initially proposed in [BSVS04] for replacing faces in arbitrary images. Temporal consistency is preserved by keeping the target face at neutral pose [JCG*08] or by applying dynamic time warping to the source sequence [DSJ*11]. Image-based methods combine robust image retrieval and image-based facial transfer techniques to select candidate source frames and warp the selected frames into the target face, respectively. To produce a seamless composite, both model-based and image-based methods blend the source into the target using either simple or tailor-made feathering operations.

Alternative image-based approaches for seamless face replacement have been proposed [BKD*08, KS16]. These methods produce compelling results, but only work on single images, i. e., no temporal constraints are imposed.

10.3. Facial Reenactment

Facial reenactment methods edit the original video face content of a target actor by transferring facial expressions from a source actor. Facial expressions are transferred by utilizing normalized motion fields [LSZ01], space-time transfer [SLS*12], motion parameters [TZN*15, TZS*16a], or even by simulating motion via image-based interpolation of candidate frames that are selected either manually [KSSGS11] or based on a similarity metric [KSSS10, LDW*14]. As a pre-requisite, these methods first reconstruct and track the source and target faces in the input videos.

One of the first reenactment methods has been proposed by Liu et al. [LSZ01]. This approach transfers both expression and shading changes to a target neutral face of an actor. Expression changes are transferred via geometric warping, while shading effects that appear in a source expression are transferred using the so-called expression ratio image, i. e., the quotient of an expressive and neutral face. The reenactment approach of Kemelmacher et al. [KSSS10] retrieves and globally aligns frames in a target video, such that the selected faces match the expression and pose of those in the source video. As the target frames are retrieved on a per-frame basis, the constructed sequence suffers from temporal jitter. Li et al. [LDW*14] proposed a temporally more coherent image-based approach. It retrieves, for each source frame, target candidates using expression and motion similarity constraints. A directed graph is then constructed and optimally traversed while enforcing selections with minimal frame jumps.

Thies et al. [TZN*15] introduced the first model-based reenactment system that runs in real time on two live RGB-D video streams (see Fig. 16). At the heart of this method is a face prior that models 3D pose, 3D face geometry, skin albedo, and incident illumination. Such a prior is used to track both the source and target performance. Fig. 17 shows that their method achieves nearly photo-realistic results. The Face2Face approach [TZS*16a] pushed the boundaries even further and proposed the first system that works on standard RGB footage, such as YouTube videos, at real-time frame rates. Unlike their initial work [TZN*15], which relies on a generic 3D teeth proxy, the Face2Face approach synthesizes the mouth interior using actual mouth frames from the target video. Fig. 18 shows that their image-based mouth synthesis produces results of higher quality.

Self-reenactment is a particular case of facial reenactment, where the source and target actor is the same person. One interesting use-case for self-reenactment was proposed in [MBW*15], where they propose to continuously blend several takes of a performance to create the desired performance in postprocess. Other examples of self-reenactment include video dubbing discussed in Section 10.5 or headset removal for VR, as proposed by FaceVR [TZS*16b] (see Fig. 19). This can, for instance, be used to allow teleconferencing in VR, where the head-mounted display (HMD) occludes half of the face of each participant.

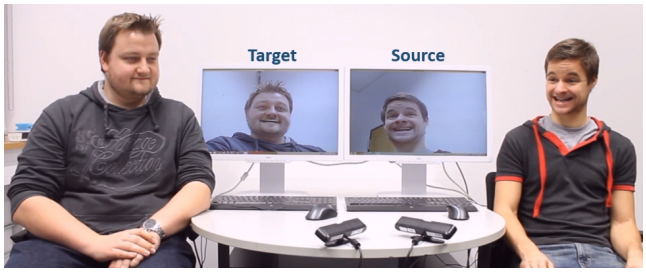

Figure 16: *Live setup used by the facial reenactment approach of Thies et al. [TZN*15]: The source actor on the right drives the target live video on the left. Both actors are captured with Asus Xtion Pro depth cameras. Based on these input observations, the faces can be reconstructed. The geometry of the target face is virtually adapted to match the expressions of the source actor. Image taken from [TZN*15].*

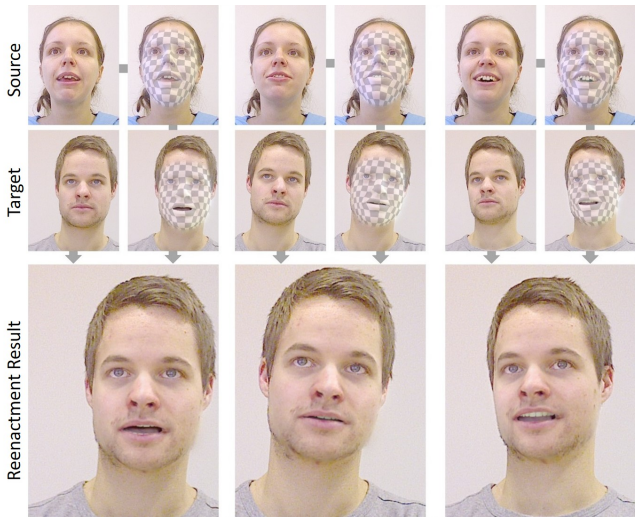

Figure 17: *RGB-D reenactment results of [TZN*15]: Based on a dense tracking energy, 3D models of both the source and target actor are reconstructed and tracked. The estimated facial expression of the source actor can be used to steer the expression of the target actor in real time, resulting in nearly photo-realistic video output. Image taken from [TZN*15].*

Recently, an image reenactment approach has been proposed [AECOKC17] that enables the creation of reactive profile pictures. In contrast to facial reenactment approaches, this approach is able to retarget slight rigid head motions based on image warping.

10.4. Speech-driven Animation

While not the focus of this report, speech-driven animation techniques have also achieved impressive results. Methods in this category typically associate units of sound, i. e., *phonemes*, to their visual counterpart, referred to as *visemes*, to synthesize new mouth animations. Phonemes are extracted either from audio streams or text using advanced speech processing tools [BHS11]. To pro-

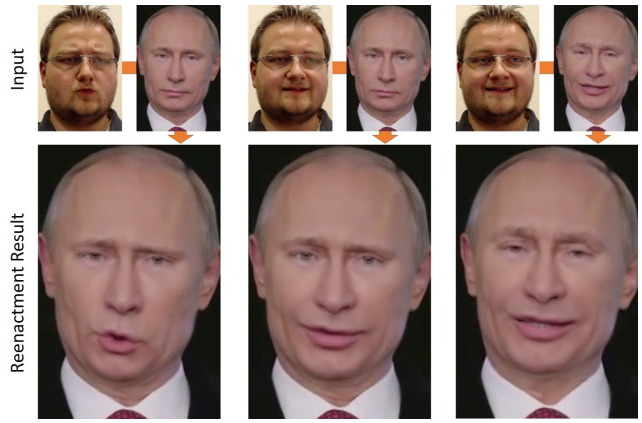

Figure 18: *Reenactment results of Face2Face [TZS*16a]: Since Face2Face reconstructs the geometry and skin reflectance of a person based on monocular RGB input, facial reenactment of Internet videos is possible. When combined with data-driven mouth synthesis, the results are indistinguishable from real photographs.*

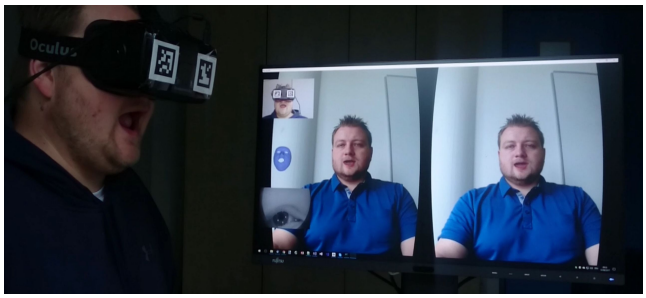

Figure 19: *FaceVR [TZS*16b] demonstrates self-reenactment for head mounted display (HMD) removal. The source actor on the left wears an HMD. His expressions are tracked using an RGB-D camera, while the eye gaze is estimated using an internal IR camera. The expression and eye gaze is then transfered to a stereoscopic video of the source actor without an HMD. This method is particularly useful for teleconference VR scenarios in which all participants can see each other's expression without occlusion. Image from [TZS*16b] (supplemental video).*

duce smooth mouth animations with proper co-articulation effects, some methods concatenate visemes based on both phoneme dissimilarity and boundary mismatch constraints [ASWC13, DN06, LO11, MCP*06, SSRMF06]. Other methods employ simple cross-fading algorithms to enforce temporal consistency between boundaries [BBPV03, BCS97, KM03]. Timing constraints can additionally be added to enhance lip sync [BCS97, TMTM12].

The output of these methods can be represented either as an animatable 3D model [DN06, KAL*17a, KM03, LXC*15b, SSRMF06, TMTM12] or as a video-realistic target video [BBPV03, CE05, EGP02, LO11]. The latter also requires solving a re-rendering and compositing problem. Recently, Suwajanakorn et al. [SSKS17] presented a deep recurrent neural network that learns a non-linear audio-to-video mapping to synthesize new person-specific mouth motions with impressive photo-realistic quality. Taylor et al. [TKY*17] presented an impressive approach based on deep learning that automatically generates natural looking speech animations of a virtual avatar based on any input speech. Similar to [SSKS17] they learn a predictor that maps phoneme label inputs to mouth movements.

The approach of Karras et al. [KAL*17b] is also able to drive facial animations based on an audio stream. They employ a deep neural network to learn a mapping from waveforms to the 3D vertex positions of the face model. The approach simultaneously computes a latent representation of facial variation that can not be explained by audio alone.

10.5. Video Dubbing

Dubbing is a post-production process used in filmmaking that adds or replaces a soundtrack in the original movie. In most cases, the original actor's voice is substituted with that of a *dubbing actor* (or *dubber*) that speaks in another language. This is a common practice in countries where subtitling is not widely accepted. Production-level dubbing requires well-trained dubbers and extensive manual interaction. A good speech-video synchronization is mandatory since viewers are very sensitive to discrepancies between the auditory and visual channel [SP54]. In fact, audio-visual mismatches can drastically impair comprehension of the spoken language [OB86, Sum92]. Unfortunately, a perfect speech-video alignment is not always feasible due to phonetic language differences.

The idea of visual dubbing is to alter the mouth motion so that it aligns with a foreign language voice spoken by the dubbing actor (see Fig. 20). Thus, visual dubbing is a particular instance of facial reenactment (see Sec. 10.3) where facial deformations are only transferred in the mouth region. In the literature, we can find speech-driven [BCS97, BBPV03, CE05, EGP02, LO11] or performance-driven [DSJ*11, TZS*16a] techniques, or even a combination of both [GVS*15]. Speech-driven dubbing techniques learn a phoneme-to-viseme mapping from a training sequence of the actor. Although these methods can produce accurate lip sync, they only work if the actor speaks both the original and foreign language. Performance-driven dubbing techniques first track the actor and dubber performance, and then transfer the dubber's mouth motion to the actor's, either using image-based swapping techniques [DSJ*11] or via parameter transfer [GVS*15, TZS*16a]. Note that the face replacement method of Dale et al. [DSJ*11] can only be employed if the actor and dubber share the same identity and lighting conditions. The VDub system of Garrido et al. [GVS*15] additionally transfers fine-scale skin detail in the mouth region and employs the audio signal to align lip closure events of salient utterances.

Fig. 21 shows that Face2Face can achieve results of similar quality to VDub for coarse-scale mouth deformations. Since Face2Face runs in real time, this method could be set up for live multilingual teleconferences where an interpreter translates a person's speech while he speaks.

Figure 20: *Visual dubbing setup of VDub [GVS*15]: The system modifies the mouth/lip motion of a target actor (a) to match a new audio track of a dubber (b+c). The dubber's mouth motion (d) is transferred to the actor, thus creating a synthesized video of the actor speaking in the dubbed language (e). Image taken from [GVS*15].*

Figure 21: *Visual dubbing results obtained by VDub [GVS*15] and Face2Face [TZS*16a]. Both methods achieve results with similar quality. Unlike VDub, Face2Face enables live dubbing. Image taken from [TZS*16a].*

10.6. Virtual Make-up

Face reconstruction and tracking at real-time rates enable compelling virtual mirror applications. Depending on the reconstruction algorithm, it is possible to change the texture of the reconstructed face, e. g., by adding virtual tattoos or logos on the face (see Fig. 22 and Fig. 23) or additionally modifying the incident illumination (Fig. 23).

Scherbaum et al. [SRH*11] developed a system that proposes the best fitting makeup style for a person, as shown in Fig. 24. They use a controlled setup to reconstruct the facial appearance information. With this information, the system learns a facial appearance

Figure 22: *Garrido et al. [GVWT13] synthesize virtual tattoos, which deform naturally with the facial expressions of the actor. Image taken from [GVWT13].*

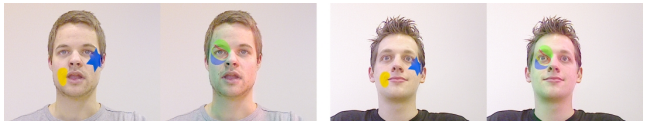

Figure 23: *Thies et al. [TZN*15] demonstrate their reconstruction quality in a virtual mirror scenario, where the skin reflectance map of a person and the illumination can be adjusted in real time. Image taken from [TZN*15].*

mapping from a face with bare skin to professional makeup. This mapping can then suggest a best-fit makeup to any new face, even when the appearance database was not trained on it.

Figure 24: *The method of Scherbaum et al. [SRH*11] suggests a facial makeup based on a learned mapping between observations with bare skin and makeup. Image taken from [SRH*11].*

Li et al. [LZL15] simulate makeup in a face image through physics-based manipulation of skin reflectance. They decompose the face image into its intrinsic image layers (albedo, diffuse shading, and specular highlights) and adapt each using advanced physically-based reflectance models that account for the interaction of the skin with cosmetics. Their method shows photo-realistic results with a multitude of appearance manipulations, i. e., various forms of cosmetics.

10.7. Projection Mapping

Projection mapping modifies the appearance of a real-world surface using projectors. FaceForge [SLS*17] demonstrates a multi-projection system that enables a user to augment a dynamically

Figure 25: *Photographs of the live projection mapping system FaceForge [SLS*17]. In this setup, two projectors are used to alter the appearance of a person in real time. Since the underlying face tracker tracks expressions, the applied new textures of the face are following the deformations of the face. Image taken from [SLS*17].*

moving and deforming human face with additional textures. In their system, two projectors are used to alter the appearance of a person. Since the underlying face tracker tracks the person's expression in real time, the applied new textures are following the deformations of the face. Fig. 25 shows photographs of live footage.

Bermano et al. [BBIG17] proposed an alternative approach that employs high-speed cameras and projectors. In combination with a simplified, yet highly efficient face tracking method, their system achieves low-latency projection mapping, also on dynamic faces.

11. Open Challenges

This report highlights the significant progress in monocular face reconstruction and tracking that has been made over the last years. Nonetheless, there are several open challenges in this field that can be tackled in the future. While face reconstruction and tracking based on controlled multi-view setups is an integral part of today's content creation pipelines and is extensively used in the creation of photo-real virtual humans, the whole pipeline is far from being fully automatized and still requires huge amounts of tedious and time-consuming manual labor. A fully automatic solution to this problem will have high impact and accelerate production times.

Current monocular reconstruction approaches made tremendous progress over the last years, but the delivered quality, accuracy and completeness of the obtained reconstructions is still far from their controlled multi-view counterparts. We think that closing this gap, especially for current real-time techniques, is of critical importance for facial virtual and augmented reality in entertainment, social media, and communication. In the following, we discuss interesting research problems in monocular face reconstruction and tracking, with a particular focus on real-time approaches. We hope that researchers in this field will tackle these problems in the future.

Most monocular reconstruction approaches heavily oversimplify the real world image formation process. A common assumption is that the illumination is distant and has low frequency. Shadows, self-shadowing and global illumination effects are ignored. While this helped in making the problem more tractable, the estimation of more accurate models will further increase the reconstruction and tracking quality, though more sophisticated priors might be needed.

Similar simplifications are often applied to the skin reflectance model. Current state-of-the-art real-time approaches default to a purely Lambertian reflectance assumption, but human faces exhibit strong specularity and a considerable amounts of subsurface scattering. These properties differ not only amongst humans, but also change dynamically based on blood flow, emotional state and sweatiness. Modeling these effects properly will lead to better reconstruction quality and enable even more convincing virtual mirror and face reenactment applications.

Most approaches to date are based on purely visual input and neglect other sources, such as audio information. Adding such modalities could improve the results substantially, in particular for performances that involve speech.

Many recent state-of-the-art monocular face reconstruction approaches only work reliably for small and medium head rotations. Larger rotations of the head lead to sever tracking failures. One reason for this is that current approaches heavily rely on sparse feature trackers that also struggle under these conditions. In the future, the robustness of face tracking under large head rotations has to be further increased to enable unconstrained face capture. We predict that deep learning based approaches will play an important role in achieving this goal.

The parts of the head that are really reconstructed by current monocular techniques, and are not just copied from a generic template, are quite restricted. Most online approaches restrict reconstruction and tracking to a tight inner face mask, where a high-quality statistical prior is readily available. Going beyond this tight mask region to a full photo-real reconstruction of a complete human head from unconstrained monocular footage is one of the major challenges for future work. Obtaining a complete reconstruction involves all the topics discussed in the advanced sections, i. e., the reconstruction of dynamic facial hair, eyes, eyelids, lips and the mouth interior. The next step goes beyond only visually complete reconstruction to the creation of complete anatomically correct rigs that obey physics, such as currently used in production. Furthermore, performances extend beyond the face and it is of paramount importance that the motion of the head is aligned to the performance of the body. This is particularly crucial in the neck area, where head and body meet, and we argue that these two should be solved jointly for optimal quality.

Deep learning methods have revolutionized computer vision over these past few years and have started to emerge also in the context of monocular face reconstruction and tracking. We expect this trend to continue and foresee hybrid solutions that combine machine learning with optimization based approaches. A crucial ingredient to such methods will be the required training data to allow for high quality capture in unconstrained scenarios. In addition to such supervised learning approaches, unsupervised approaches seem to be promising and might be able to simplify the expensive and tedious creation of the employed training corpora.

Recent publications like Face2Face [TZS*16a] and Synthesizing Obama [SSKS17] have also raised questions about ethical and safety issues. With the progress of face reconstruction and photo-realistic synthesis, we expect these questions to come more and more to the forefront, and addressing these issues openly within the scientific community will become more and more important.

12. Conclusion

In this report, we have provided an overview of recent trends in cutting-edge research for monocular facial performance capture, and we have discussed current applications in performance-based facial animation, video manipulation, and real-time facial reenactment. Based on the foundations of real-world image formation, we have shown how deformable parametric face models can be robustly fitted to monocular RGB and RGB-D video, even at real-time frame rates. Furthermore, we have covered high-quality multi-view techniques that are the foundation to build better priors and go beyond the confined face mask, which currently forms the basis in most state-of-the-art approaches.

We hope that this survey serves as a comprehensive guide to this

fascinating and active field, and that our discussion of open challenges will motivate important follow-up work, thus further narrowing the gap between monocular and high-quality multi-view approaches. We are also convinced that these trends will further increase the availability of monocular face reconstruction and tracking technology, and will have a high impact on the development of facial virtual and augmented reality applications. We believe that in the future real-time high-quality facial capture will play a central role in digital communication, and commodity devices will enable mass deployment of facial reconstruction and tracking techniques, a trend that has already begun with the latest release of Apple's iPhone X and their *Face ID* unlocking system.

Acknowledgements

This work was supported by the ERC Starting Grant CapReal (335545), the Max Planck Center for Visual Computing and Communications (MPC-VCC), a TUM-IAS Rudolf Mößbauer Fellowship and a Google Faculty Award.

References

[AECOKC17] AVERBUCH-ELOR H., COHEN-OR D., KOPF J., COHEN M. F.: Bringing portraits to life. *ACM Trans. Graph. 36*, 6 (2017), 196:1–196:13. 16

[AFB*13] ALEXANDER O., FYFFE G., BUSCH J., YU X., ICHIKARI R., JONES A., DEBEVEC P., JIMENEZ J., DANVOYE E., ANTIONAZZI B., EHELER M., KYSELA Z., VON DER PAHLEN J.: Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters* (New York, NY, USA, 2013), SIGGRAPH '13, ACM, pp. 1:1–1:1. 3, 14, 27

[ANRS07] ABATE A. F., NAPPI M., RICCIO D., SABATINO G.: 2d and 3d face recognition: A survey. *Pattern Recogn. Lett. 28*, 14 (2007), 1885–1906. 2

[ARL*10] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG J.-Y., MA W.-C., WANG C.-C., DEBEVEC P.: The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications 30*, 4 (2010), 20–31. 3, 14

[ASC13] ANDERSON R., STENGER B., CIPOLLA R.: Lip tracking for 3d face registration. In *Proceedings of the 13. IAPR International Conference on Machine Vision Applications* (2013), MVA '13, pp. 145–148. 12

[ASWC13] ANDERSON R., STENGER B., WAN V., CIPOLLA R.: Expressive visual text-to-speech using active appearance models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), CVPR '13, IEEE Computer Society, pp. 3382–3389. 17

[AZCP13] ASTHANA A., ZAFEIRIOU S., CHENG S., PANTIC M.: Robust discriminative response map fitting with constrained local models. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2013), CVPR '13, IEEE Computer Society, pp. 3444–3451. 9, 26

[BB01] BRAND M., BHOTIKA R.: Flexible flow for 3d nonrigid tracking and shape recovery. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition* (2001), CVPR '01, IEEE Computer Society, pp. 315–322. 26

[BB14] BEELER T., BRADLEY D.: Rigid stabilization of facial expressions. *ACM Trans. Graph. 33*, 4 (2014), 44:1–44:9. 8

[BBA*07] BICKEL B., BOTSCH M., ANGST R., MATUSIK W., OTADUY M., PFISTER H., GROSS M.: Multi-scale capture of facial geometry and motion. *ACM Trans. Graph. 26*, 3 (2007), 33:1–33:10. 3, 7

[BBB*14] BERMANO A. H., BRADLEY D., BEELER T., ZUND F., NOWROUZEZAHRAI D., BARAN I., SORKINE-HORNUNG O., PFISTER H., SUMNER R. W., BICKEL B., GROSS M.: Facial performance enhancement using dynamic shape space analysis. *ACM Trans. Graph. 33*, 2 (2014), 13:1–13:12. 7, 15

[BBGB16] BÉRARD P., BRADLEY D., GROSS M., BEELER T.: Lightweight eye capture using a parametric model. *ACM Trans. Graph. 35*, 4 (2016), 117:1–117:12. 13

[BBIG17] BERMANO A. H., BILLETER M., IWAI D., GRUNDHÖFER A.: Makeup lamps: Live augmentation of human faces via projection. *Comput. Graph. Forum 36*, 2 (2017), 311–323. 19

[BBK*15] BERMANO A., BEELER T., KOZLOV Y., BRADLEY D., BICKEL B., GROSS M.: Detailed spatio-temporal reconstruction of eyelids. *ACM Trans. Graph. 34*, 4 (2015), 44:1–44:11. 13

[BBN*12] BEELER T., BICKEL B., NORIS G., BEARDSLEY P., MARSCHNER S., SUMNER R. W., GROSS M.: Coupled 3d reconstruction of sparse facial hair and skin. *ACM Trans. Graph. 31*, 4 (2012), 117:1–117:10. 14

[BBN*14] BÉRARD P., BRADLEY D., NITTI M., BEELER T., GROSS M.: High-quality capture of eyes. *ACM Trans. Graph. 33*, 6 (2014), 223:1–223:12. 13

[BBPV03] BLANZ V., BASSO C., POGGIO T., VETTER T.: Reanimating faces in images and video. *Comput. Graph. Forum 22* (2003), 641–650. 17, 26

[BBW14] BRUNTON A., BOLKART T., WUHRER S.: Multilinear wavelets: A statistical shape space for human faces. In *European Conference on Computer Vision* (2014), Springer, pp. 297–312. 6

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1997), SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., pp. 353–360. 17

[BGY*13] BHAT K. S., GOLDENTHAL R., YE Y., MALLET R., KOPERWAS M.: High fidelity facial animation capture and retargeting with contours. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2013), SCA '13, ACM, pp. 7–14. 12

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph. 30*, 4 (2011), 75:1–75:10. 3, 7, 12

[BHPS10] BRADLEY D., HEIDRICH W., POPA T., SHEFFER A.: High resolution passive facial performance capture. *ACM Trans. Graph. 29*, 4 (2010), 41:1–41:10. 3, 7, 12

[BHS11] BERGER M. A., HOFER G., SHIMODAIRA H.: Carnival–combining speech technology and computer animation. *IEEE Computer Graphics and Applications 31*, 5 (2011), 80–89. 16

[Bjö96] BJÖRCK A.: *Numerical Methods for Least Squares Problems*. Siam Philadelphia, 1996. 11

[BK13] BRADSKI G., KAEHLER A.: *Learning OpenCV: Computer Vision in C++ with the OpenCV Library*, 2nd ed. O'Reilly Media, Inc., 2013. 26

[BKD*08] BITOUK D., KUMAR N., DHILLON S., BELHUMEUR P., NAYAR S. K.: Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph. 27*, 3 (2008), 39:1–39:8. 16

[BL03] BORSHUKOV G., LEWIS J. P.: Realistic human face rendering for "the matrix reloaded". In *ACM SIGGRAPH 2003 Sketches & Applications* (New York, NY, USA, 2003), SIGGRAPH '03, ACM, pp. 16:1–16:1. 3, 4

[BLB*08] BICKEL B., LANG M., BOTSCH M., OTADUY M. A., GROSS M.: Pose-space animation and transfer of facial details. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2008), SCA '08, Eurographics Association, pp. 57–66. 7

[BRM12] BALTRUŠAITIS T., ROBINSON P., MORENCY L.-P.: 3d constrained local model for rigid and non-rigid facial tracking. In *IEEE Conference on Computer Vision and Pattern Recognition* (2012). 9

[BRZ*16] BOOTH J., ROUSSOS A., ZAFEIRIOU S., PONNIAH A., DUNAWAY D.: A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016). 5

[BSVS04] BLANZ V., SCHERBAUM K., VETTER T., SEIDEL H.: Exchanging faces in images. *Comput. Graph. Forum 23*, 3 (2004), 669–676. 15

[BTLP16] BOUAZIZ S., TAGLIASACCHI A., LI H., PAULY M.: Modern techniques and applications for real-time non-rigid registration. In *SIGGRAPH ASIA 2016 Courses* (New York, NY, USA, 2016), SA '16, ACM, pp. 11:1–11:25. 2

[BTP14] BOUAZIZ S., TAGLIASACCHI A., PAULY M.: Dynamic 2d/3d registration. 2

[BV99] BLANZ V., VETTER T.: A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1999), SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., pp. 187–194. 5, 6, 9, 10

[BWP13] BOUAZIZ S., WANG Y., PAULY M.: Online modeling for realtime facial animation. *ACM Trans. Graph. 32*, 4 (2013), 40:1–40:10. 3, 4, 5, 7, 8, 9, 10, 12, 15, 26

[BY95] BLACK M. J., YACOOB Y.: Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the 5th International Conference on Computer Vision* (1995), ICCV '95, pp. 374–381. 26

[CBB*15] CONG M., BAO M., BHAT K. S., FEDKIW R., ET AL.: Fully automatic generation of anatomical face simulation models. In *Proceedings of the 14th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2015), ACM, pp. 175–183. 8

[CBZB15] CAO C., BRADLEY D., ZHOU K., BEELER T.: Real-time high-fidelity facial performance capture. *ACM Trans. Graph. 34*, 4 (2015), 46:1–46:9. 3, 4, 7, 9, 26

[CC06] CRISTINACCE D., COOTES T. F.: Feature detection and tracking with constrained local models. In *Proceedings of the 2006 British Machine Vision Conference* (2006), BMVC '06, British Machine Vision Association, pp. 929–938. 9, 26

[CE05] CHANG Y.-J., EZZAT T.: Transferable videorealistic speech animation. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2005), SCA '05, ACM, pp. 143–151. 17

[CET01] COOTES T. F., EDWARDS G. J., TAYLOR C. J.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell. 23*, 6 (2001), 681–685. 9, 26

[CFKP04] CAO Y., FALOUTSOS P., KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2004), SCA '04, Eurographics Association, pp. 345–353. 5

[CHZ14] CAO C., HOU Q., ZHOU K.: Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph. 33*, 4 (2014), 43:1–43:10. 3, 4, 8, 10, 11, 15, 26

[CLS*15] CHAI M., LUO L., SUNKAVALLI K., CARR N., HADAP S., ZHOU K.: High-quality hair modeling from a single portrait photo. *ACM Trans. Graph. 34*, 6 (2015), 204:1–204:10. 14

[CM92] CHEN Y., MEDIONI G. G.: Object modelling by registration of multiple range images. *Image and Vision Computing 10*, 3 (1992), 145–155. 10

[CSW*16] CHAI M., SHAO T., WU H., WENG Y., ZHOU K.: Autohair: Fully automatic hair modeling from a single image. *ACM Trans. Graph. 35*, 4 (2016), 116:1–116:12. 14

[CTCG95] COOTES T. F., TAYLOR C. J., COOPER D. H., GRAHAM J.: Active shape models—their training and application. *Comput. Vis. Image Underst. 61*, 1 (1995), 38–59. 9

[CWLZ13] CAO C., WENG Y., LIN S., ZHOU K.: 3D shape regression for real-time facial animation. *ACM Trans. Graph. 32*, 4 (2013), 41:1–41:10. 3, 4, 8, 10, 11, 15, 26

[CWS*13] CHEN Y.-L., WU H.-T., SHI F., TONG X., CHAI J.: Accurate and robust 3d facial capture using a single rgbd camera. *2013 IEEE International Conference on Computer Vision* (2013), 3615–3622. 3, 8, 9, 10, 12

[CWW*16] CAO C., WU H., WENG Y., SHAO T., ZHOU K.: Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph. 35*, 4 (2016), 126:1–126:12. 9, 12, 14, 15

[CWZ*14] CAO C., WENG Y., ZHOU S., TONG Y., ZHOU K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2014), 413–425. 5

[CXH03] CHAI J.-X., XIAO J., HODGINS J.: Vision-based control of 3d facial animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2003), SCA '03, Eurographics Association, pp. 193–206. 15

[CZZ14] CHAI M., ZHENG C., ZHOU K.: A reduced model for interactive hairs. *ACM Trans. Graph. 33*, 4 (2014), 124:1–124:11. 14

[DI4] DI4D: Dimensional imaging. http://www.di4d.com/. 3

[DJBDDT13] DEROUET-JOURDAN A., BERTAILS-DESCOUBES F., DAVIET G., THOLLOT J.: Inverse dynamic hair modeling with frictional contact. *ACM Trans. Graph. 32*, 6 (2013). 14

[DM96] DECARLO D., METAXAS D.: The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *Proceedings of the 1996 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 1996), CVPR '96, IEEE Computer Society, pp. 231–238. 26

[DM00] DECARLO D., METAXAS D.: Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision 38*, 2 (2000), 99–127. 6

[DN06] DENG Z., NEUMANN U.: efase: Expressive facial animation synthesis and editing with phoneme-isomap controls. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2006), SCA '06, Eurographics Association, pp. 251–260. 17

[DSJ*11] DALE K., SUNKAVALLI K., JOHNSON M. K., VLASIC D., MATUSIK W., PFISTER H.: Video face replacement. *ACM Trans. Graph. 30*, 6 (2011), 130:1–130:10. 15, 17

[DSVG10] DE SMET M., VAN GOOL L.: Optimal regions for linear model-based 3d face reconstruction. In *Asian Conference on Computer Vision* (2010), Springer, pp. 276–289. 6

[EBDP96] ESSA I., BASU S., DARRELL T., PENTLAND A.: Modeling, tracking and interactive animation of faces and heads using input from video. In *Proceedings of the Computer Animation* (Washington, DC, USA, 1996), CA '96, IEEE Computer Society, pp. 68–79. 7, 26

[EBGB14] ECHEVARRIA J. I., BRADLEY D., GUTIERREZ D., BEELER T.: Capturing and stylizing hair for 3d fabrication. *ACM Trans. Graph. 33*, 4 (2014), 125:1–125:11. 14

[EG98] EISERT P., GIROD B.: Analyzing facial expressions for virtual conferencing. *IEEE Comput. Graph. Appl. 18*, 5 (1998), 70–78. 15

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. *ACM Trans. Graph. 21*, 3 (2002), 388–398. 17

[FJA*14] FYFFE G., JONES A., ALEXANDER O., ICHIKARI R., DEBEVEC P.: Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph. 34*, 1 (2014), 8:1–8:14. 3, 7

[FP09] FURUKAWA Y., PONCE J.: Dense 3D motion capture for human faces. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), CVPR '09, IEEE Computer Society, pp. 1674–1681. 3, 7

[FP12] FORSYTH D. A., PONCE J.: *Computer Vision: A Modern Approach*, 2nd ed. Pearson, 2012. 5

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph. 30*, 6 (2011), 129:1–129:10. 3

[GGG*16] GUARNERA D., GUARNERA G., GHOSH A., DENK C., GLENCROSS M.: Brdf representation and acquisition. *Computer Graphics Forum 35*, 2 (2016), 625–650. 26

[GGW*98] GUENTER B., GRIMM C., WOOD D., MALVAR H., PIGHIN F.: Making faces. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1998), SIGGRAPH '98, ACM, pp. 55–66. 1

[GRA13] GARG R., ROUSSOS A., AGAPITO L.: A variational approach to video registration with subspace constraints. *International Journal of Computer Vision 104*, 3 (2013), 286–314. 26

[GVR*14] GARRIDO P., VALGAERTS L., REHMSEN O., THORMAEHLEN T., PÉREZ P., THEOBALT C.: Automatic face reenactment. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2014), CVPR '14, IEEE Computer Society, pp. 4217–4224. 15

[GVS*15] GARRIDO P., VALGAERTS L., SARMADI H., STEINER I., VARANASI K., PÉREZ P., THEOBALT C.: VDub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum 34*, 2 (2015), 193–204. 13, 15, 17, 18

[GVWT13] GARRIDO P., VALGAERTS L., WU C., THEOBALT C.: Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph. (Proceedings of SIGGRAPH Asia 2013) 32*, 6 (2013), 158:1–158:10. 3, 4, 7, 8, 9, 18, 27

[GZC*16] GARRIDO P., ZOLLHÖFER M., CASAS D., VALGAERTS L., VARANASI K., PÉREZ P., THEOBALT C.: Reconstruction of personalized 3d face rigs from monocular video. *ACM Trans. Graph. 35*, 3 (2016), 28:1–28:15. 3, 4, 5, 7, 8, 9, 10, 12, 15, 26, 27

[GZW*16] GARRIDO P., ZOLLHÖFER M., WU C., BRADLEY D., PÉREZ P., BEELER T., THEOBALT C.: Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph. 35*, 6 (2016), 219:1–219:11. 6, 12

[HBLB17] HU L., BRADLEY D., LI H., BEELER T.: Simulation-ready hair capture. *Computer Graphics Forum 36*, 2 (2017), 281–294. 14

[HCTW11] HUANG H., CHAI J., TONG X., WU H.-T.: Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. *ACM Trans. Graph. 30*, 4 (2011), 74:1–74:10. 3

[HJ10] HANSEN D. W., JI Q.: In the eye of the beholder: A survey of models for eyes and gaze. *IEEE PAMI 32*, 3 (2010), 478–500. 13

[HML*14] HU L., MA C., LUO L., WEI L.-Y., LI H.: Capturing braided hairstyles. *ACM Trans. Graph. 33*, 6 (2014), 225:1–225:9. 14

[HMLL14] HU L., MA C., LUO L., LI H.: Robust hair capture using simulated examples. *ACM Trans. Graph. 33*, 4 (2014). 14

[HMLL15] HU L., MA C., LUO L., LI H.: Single-view hair modeling using a hairstyle database. *ACM Trans. Graph. 34*, 4 (2015), 125:1–125:9. 14

[HMYL15] HSIEH P., MA C., YU J., LI H.: Unconstrained realtime facial performance capture. In *Poceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* (2015), CVPR '15, IEEE Computer Society, pp. 1675–1683. 3, 4, 9, 10, 12

[Hor75] HORN B. K.: Obtaining shape from shading information. *The psychology of computer vision* (1975), 115–155. 7

[HSB*16] HEWER A., STEINER I., BOLKART T., WUHRER S., RICHMOND K.: A statistical shape space model of the palate surface trained on 3d MRI scans of the vocal tract. *CoRR abs/1602.07679* (2016). 13

[HSW14] HEWER A., STEINER I., WUHRER S.: A hybrid approach to 3d tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association* (2014), INTERSPEECH '14, ISCA, pp. 418–421. 13

[HSW*17] HU L., SAITO S., WEI L., NAGANO K., SEO J., FURSUND J., SADEGHI I., SUN C., CHEN Y., LI H.: Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph. 36*, 6 (2017), 195:1–195:14. 14

[HYZ*12] HUANG H., YIN K., ZHAO L., QI Y., YU Y., TONG X.: Detail-preserving controllable deformation from sparse examples. *IEEE Transactions on Visualization and Computer Graphics 18*, 8 (2012), 1215–1227. 7

[HZW*04] HUANG X., ZHANG S., WANG Y., METAXAS D. N., SAMARAS D.: A hierarchical framework for high resolution facial expression tracking. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Washington, DC, USA, 2004), CVPR Workshops '04, IEEE Computer Society, p. 22. 3

[HZW12] HERRERA T. L., ZINKE A., WEBER A.: Lighting hair from the inside: A thermal approach to hair reconstruction. *ACM Trans. Graph. 31*, 6 (2012), 146:1–146:9. 14

[IBP15] ICHIM A. E., BOUAZIZ S., PAULY M.: Dynamic 3d avatar creation from hand-held video input. *ACM Trans. Graph. 34*, 4 (2015), 45:1–45:14. 4, 5, 6, 7, 8, 10, 12, 14, 15, 26

[IKKP17] ICHIM A.-E., KADLEČEK P., KAVAN L., PAULY M.: Phace: Physics-based face modeling and animation. *ACM Trans. Graph. 36*, 4 (July 2017), 153:1–153:14. 8

[IM] IM: Image metrics. http://www.image-metrics.com/. 3

[JCG*08] JONES A., CHIANG J., GHOSH A., LANG M., HULLIN M., BUSCH J., DEBEVEC P.: *Real-time Geometry and Reflectance Capture for Digital Face Replacement*. Tech. Rep. 4s, University of Southern California, 2008. 15

[JMLH01] JENSEN H. W., MARSCHNER S. R., LEVOY M., HANRAHAN P.: A practical model for subsurface light transport. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, ACM, pp. 511–518. 4

[JTDP03] JOSHI P., TIEN W. C., DESBRUN M., PIGHIN F.: Learning controls for blend shape based realistic facial animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2003), SCA '03, Eurographics Association, pp. 187–192. 6

[KAL*17a] KARRAS T., AILA T., LAINE S., HERVA A., LEHTINEN J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. 36*, 4 (July 2017), 94:1–94:12. 17

[KAL*17b] KARRAS T., AILA T., LAINE S., HERVA A., LEHTINEN J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. 36*, 4 (2017), 94:1–94:12. 17

[KH12] KLAUDINY M., HILTON A.: High-detail 3d capture and non-sequential alignment of facial performance. In *Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission* (2012), 3DIMPVT '12, IEEE Computer Society, pp. 17–24. 3

[KM03] KSHIRSAGAR S., MAGNENAT-THALMANN N.: Visyllable based speech animation. *Comput. Graph. Forum 22*, 3 (2003), 632–640. 17

[KMB*17] KLAUDINY M., MCDONAGH S., BRADLEY D., BEELER T., MITCHELL K.: Real-Time Multi-View Facial Capture with Synthetic Training. *Computer Graphics Forum* (2017). 11, 12

[KMML10] KOMOROWSKI D., MELAPUDI V., MORTILLARO D., LEE G. S.: A hybrid approach to facial rigging. In *ACM SIGGRAPH ASIA 2010 Sketches* (New York, NY, USA, 2010), SA '10, ACM, pp. 42:1–42:2. 7

[KRP*15] KLEHM O., ROUSSELLE F., PAPAS M., BRADLEY D., HERY C., BICKEL B., JAROSZ W., BEELER T.: Recent advances in facial appearance capture. *Comput. Graph. Forum 34*, 2 (2015), 709–733. 2, 4, 27

[KS13] KEMELMACHER-SHLIZERMAN I.: Internet-based morphable model. *International Conference on Computer Vision (ICCV)* (2013). 6

[KS16] KEMELMACHER-SHLIZERMAN I.: Transfiguring portraits. *ACM Trans. Graph. 35*, 4 (July 2016), 94:1–94:8. 16

[KSSGS11] KEMELMACHER SHLIZERMAN I., SHECHTMAN E., GARG R., SEITZ S. M.: Exploring photobios. *ACM Trans. Graph. 30*, 4 (2011), 61:1–61:10. 16

[KSSS10] KEMELMACHER-SHLIZERMAN I., SANKAR A., SHECHTMAN E., SEITZ S. M.: Being john malkovich. In *Proceedings of the 11th European Conference on Computer Vision* (2010), vol. 6311 of *Lecture Notes in Computer Science*, Springer, pp. 341–353. 16

[LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F., DENG Z.: Practice and theory of blendshape facial models. In *Eurographics 2014 - State of the Art Reports* (2014), Lefebvre S., Spagnuolo M., (Eds.), The Eurographics Association, pp. 199–218. 2, 5

[Law07] LAWRENCE N. D.: Learning for larger datasets with the gaussian process latent variable model. In *Artificial Intelligence and Statistics* (2007), pp. 243–250. 8

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph. 36*, 6 (Nov. 2017), 194:1–194:17. 14

[LDW*14] LI K., DAI Q., WANG R., LIU Y., XU F., WANG J.: A data-driven approach for facial expression retargeting in video. *IEEE Trans. Multimedia 16*, 2 (2014), 299–310. 16

[Lev44] LEVENBERG K.: A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics 2*, 2 (1944), 164–168. 11

[LKA*17] LAINE S., KARRAS T., AILA T., HERVA A., SAITO S., YU R., LI H., LEHTINEN J.: Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation* (New York, NY, USA, 2017), SCA '17, ACM, pp. 10:1–10:10. 11, 12

[LLP*12] LUO L., LI H., PARIS S., WEISE T., PAULY M., RUSINKIEWICZ S.: Multi-view hair capture using orientation fields. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), CVPR '12, IEEE Computer Society, pp. 1490–1497. 14

[LO11] LIU K., OSTERMANN J.: Realistic facial expression synthesis for an image-based talking head. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo* (2011), ICME '11, IEEE Computer Society, pp. 1–6. 17

[LRF93] LI H., ROIVAINEN P., FORCHEIMER R.: 3dd motion estimation in model-based facial image coding. *IEEE Trans. Pattern Anal. Mach. Intell. 15*, 6 (1993), 545–555. 26

[LSKS16] LIANG S., SHAPIRO L. G., KEMELMACHER-SHLIZERMAN I.: Head reconstruction from internet photos. In *European Conference on Computer Vision* (2016), Springer, pp. 360–374. 14

[LSZ01] LIU Z., SHAN Y., ZHANG Z.: Expressive expression mapping with ratio images. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, ACM, pp. 271–276. 16

[LWP10] LI H., WEISE T., PAULY M.: Example-based facial rigging. *ACM Trans. Graph. 29*, 4 (2010), 32:1–32:6. 5, 7, 8

[LXC*15a] LI J., XU W., CHENG Z., XU K., KLEIN R.: Lightweight wrinkle synthesis for 3d facial modeling and animation. *Computer-Aided Design 58* (2015), 117–122. 7

[LXC*15b] LIU Y., XU F., CHAI J., TONG X., WANG L., HUO Q.: Video-audio driven real-time facial animation. *ACM Trans. Graph. 34*, 6 (2015), 182:1–182:10. 17

[LYYB13] LI H., YU J., YE Y., BREGLER C.: Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph. 32*, 4 (2013), 42:1–42:10. 3, 4, 5, 7, 8, 9, 10, 12, 15, 26

[LZL15] LI C., ZHOU K., LIN S.: Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* (2015), CVPR '15, IEEE Computer Society, pp. 4621–4629. 18

[Mar63] MARQUARDT D. W.: An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics 11*, 2 (1963), 431–441. 11

[MB04] MATTHEWS I., BAKER S.: Active appearance models revisited. *Int. J. Comput. Vision 60*, 2 (2004), 135–164. 9, 26

[MBW*15] MALLESON C., BAZIN J.-C., WANG O., BRADLEY D., BEELER T., HILTON A., SORKINE-HORNUNG A.: Facedirector: continuous control of facial performance in video. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3979–3987. 16

[MCP*06] MA J., COLE R. A., PELLOM B. L., WARD W. H., WISE B.: Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Trans. Vis. Comput. Graph. 12*, 2 (2006), 266–276. 17

[MHP*07] MA W.-C., HAWKINS T., PEERS P., CHABERT C.-F., WEISS M., DEBEVEC P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Aire-la-Ville, Switzerland, Switzerland, 2007), EGSR '07, Eurographics Association, pp. 183–194. 3

[MJC*08] MA W.-C., JONES A., CHIANG J.-Y., HAWKINS T., FREDERIKSEN S., PEERS P., VUKOVIC M., OUHYOUNG M., DEBEVEC P.: Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Trans. Graph. 27*, 5 (2008), 121:1–121:10. 7

[MKB*16] MCDONAGH S., KLAUDINY M., BRADLEY D., BEELER T., MATTHEWS I., MITCHELL K.: Synthetic prior design for real-time face tracking. *2016 Fourth International Conference on 3D Vision (3DV) 00* (2016), 639–648. 11, 12

[MM76] MCGURK H., MACDONALD J.: Hearing lips and seeing voices. *Nature 264*, 5588 (1976), 746–748. 12

[Mor78] MORÉ J. J.: *The Levenberg-Marquardt algorithm: Implementation and theory*, vol. 630 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1978, pp. 105–116. 11

[MOV] MOVA: Mova®contour®facial capture system. http://rearden.com/mova.html. 3

[Mue66] MUELLER C.: *Spherical harmonics*, vol. 17 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 1966. 27

[NN01] NOH J.-Y., NEUMANN U.: Expression cloning. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, ACM, pp. 277–288. 15

[NVW*13] NEUMANN T., VARANASI K., WENGER S., WACKER M., MAGNOR M., THEOBALT C.: Sparse localized deformation components. *ACM Trans. Graph. 32*, 6 (2013), 179:1–179:10. 6, 7

[NW06] NOCEDAL J., WRIGHT S. J.: *Numerical Optimization*, 2nd ed. Springer, New York, 2006. 11

[OB86] OWENS E., BLAZEK B.: Visemes observed by the hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research 28* (1986), 381–393. 17

[OBP*12] ORVALHO V., BASTOS P., PARKE F., OLIVEIRA B., ALVAREZ X.: A facial rigging survey. In *Eurographics State of The Art Reports* (2012). 2

[OLSL16] OLSZEWSKI K., LIM J. J., SAITO S., LI H.: High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2016) 35*, 6 (December 2016). 11, 12

[PCK*08] PARIS S., CHANG W., KOZHUSHNYAN O. I., JAROSZ W., MATUSIK W., ZWICKER M., DURAND F.: Hair photobooth: Geometric and photometric acquisition of real hairstyles. *ACM Trans. Graph. 27*, 3 (2008). 14

[PL06] PIGHIN F., LEWIS J.: Performance-driven facial animation. In *ACM SIGGRAPH Courses* (2006). 3

[PSS99] PIGHIN F., SZELISKI R., SALESIN D.: Resynthesizing facial animation through 3D model-based tracking. In *Proceedings of the 7th International Conference on Computer Vision* (1999), ICCV '99, IEEE Computer Society, pp. 143–150. 26

[RH01a] RAMAMOORTHI R., HANRAHAN P.: An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, ACM, pp. 497–500. 28

[RH01b] RAMAMOORTHI R., HANRAHAN P.: A signal-processing framework for inverse rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2001), SIGGRAPH '01, ACM, pp. 117–128. 28

[RSK16] RICHARDSON E., SELA M., KIMMEL R.: 3d face reconstruction by learning from synthetic data. In *Proceedings of the Fourth International Conference on 3D Vision* (2016), 3DV '16, IEEE Computer Society, pp. 460–469. 11, 12

[RSOK17] RICHARDSON E., SELA M., OR-EL R., KIMMEL R.: Learning detailed face reconstruction from a single image. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (2017), CVPR '17, IEEE Computer Society, pp. 5553–5562. 11, 12

[RZL*17] RIBERA R. B. I., ZELL E., LEWIS J. P., NOH J., BOTSCH M.: Facial retargeting with automatic range of motion alignment. *ACM Trans. Graph. 36*, 4 (2017), 154:1–154:12. 15

[SDT*07] SONG M., DONG Z., THEOBALT C., WANG H., LIU Z., SEIDEL H. P.: A generic framework for efficient 2-d and 3-d facial expression analogy. *IEEE Trans. Multimedia 9*, 7 (2007), 1384–1395. 15

[SEL17] SEYMOUR M., EVANS C., LIBRERI K.: Meet mike: Epic avatars. In *ACM SIGGRAPH 2017 VR Village* (New York, NY, USA, 2017), SIGGRAPH '17, ACM, pp. 12:1–12:2. 3, 14

[SL09] STYLIANOU G., LANITIS A.: Image based 3d face reconstruction: A survey. *International Journal of Image and Graphics 09*, 02 (2009), 217–250. 2

[SLC11a] SARAGIH J. M., LUCEY S., COHN J. F.: Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vision 91*, 2 (2011), 200–215. 9, 15, 26

[SLC11b] SARAGIH J. M., LUCEY S., COHN J. F.: Real-time avatar animation from a single image. In *Proceedings of the Ninth IEEE International Conference on Automatic Face and Gesture Recognition* (2011), FG '11, IEEE Computer Society, pp. 117–124. 15

[SLL16] SAITO S., LI T., LI H.: Real-time facial segmentation and performance capture from RGB input. In *Proceedings of the 14th European Conference on Computer Vision* (2016), vol. 9912 of *Lecture Notes in Computer Science*, Springer, pp. 244–261. 3, 4, 9, 10, 12, 26

[SLS*12] SEOL Y., LEWIS J., SEO J., CHOI B., ANJYO K., NOH J.: Spacetime expression cloning for blendshapes. *ACM Trans. Graph. 31*, 2 (2012), 14:1–14:12. 16

[SLS*17] SIEGL C., LANGE V., STAMMINGER M., BAUER F., THIES J.: Faceforge: Markerless non-rigid face multi-projection mapping. *IEEE Transactions on Visualization and Computer Graphics* (2017). 18

[SNF05] SIFAKIS E., NEVEROV I., FEDKIW R.: Automatic determination of facial muscle activations from sparse motion capture marker data. In *Acm transactions on graphics (tog)* (2005), vol. 24, ACM, pp. 417–425. 8

[SP54] SUMBY W. H., POLLACK I.: Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America 26*, 2 (1954), 212–215. 17

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. *ACM Trans. Graph. 23*, 3 (2004), 399–405. 5

[SRH*11] SCHERBAUM K., RITSCHEL T., HULLIN M., THORMÃĎHLEN T., BLANZ V., SEIDEL H.-P.: Computer-Suggested Facial Makeup. *Computer Graphics Forum* (2011). 18

[SRK17] SELA M., RICHARDSON E., KIMMEL R.: Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision* (2017), ICCV '17, IEEE Computer Society, pp. 1585–1594. 11, 12

[SRT*11] STURM P., RAMALINGAM S., TARDIF J.-P., GASPARINI S., BARRETO J. A.: Camera models and fundamental concepts used in geometric computer vision. *Found. Trends. Comput. Graph. Vis. 6* (Jan. 2011), 1–183. 5

[SSK15] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: What makes tom hanks look like tom hanks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision* (2015), ICCV '15, IEEE Computer Society, pp. 3952–3960. 15

[SSKS17] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. 36*, 4 (July 2017), 95:1–95:13. 17, 19

[SSRMF06] SIFAKIS E., SELLE A., ROBINSON-MOSHER A., FEDKIW R.: Simulating speech with a physics-based facial muscle model. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2006), SCA '06, Eurographics Association, pp. 261–270. 7, 17

[SSS14] SUWAJANAKORN S., SHLIZERMAN I. K., SEITZ S. M.: Total moving face reconstruction. In *Proceedings of the 13th European Conference on Computer Vision* (2014), vol. 8692 of *Lecture Notes in Computer Science*, Springer, pp. 796–812. 3, 4, 7, 8, 9, 10, 12, 26, 27

[Sum92] SUMMERFIELD Q.: Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 335*, 1273 (1992), 71–78. 17

[SWTC14] SHI F., WU H.-T., TONG X., CHAI J.: Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph. 33*, 6 (2014), 222:1–222:13. 3, 4, 7, 8, 9, 10, 12, 26

[SZPY12] SANDBACH G., ZAFEIRIOU S., PANTIC M., YIN L.: Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image Vision Comput. 30*, 10 (2012), 683–697. 2

[TDlTM11] TENA J. R., DE LA TORRE F., MATTHEWS I.: Interactive region-based linear 3d face models. *ACM Trans. Graph. 30*, 4 (2011), 76:1–76:10. 6

[TKY*17] TAYLOR S., KIM T., YUE Y., MAHLER M., KRAHE J., RODRIGUEZ A. G., HODGINS J., MATTHEWS I.: A deep learning approach for generalized speech animation. *ACM Trans. Graph. 36*, 4 (July 2017), 93:1–93:11. 17

[TMM*09] THEOBALD B.-J., MATTHEWS I., MANGINI M., SPIES J. R., BRICK T. R., COHN J. F., BOKER S. M.: Mapping and manipulating facial expression. *Language and Speech 52*, 2–3 (2009), 369–386. 15

[TMTM12] TAYLOR S. L., MAHLER M., THEOBALD B.-J., MATTHEWS I.: Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2012), SCA '12, Eurographics Association, pp. 275–284. 17

[TTHMM17] TUAN TRAN A., HASSNER T., MASI I., MEDIONI G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017). 11, 12

[TZK*17] TEWARI A., ZOLLÖFER M., KIM H., GARRIDO P., BERNARD F., PEREZ P., CHRISTIAN T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proc. IEEE International Conference on Computer Vision (ICCV)* (2017). 11, 12

[TZN*15] THIES J., ZOLLHÖFER M., NIESSNER M., VALGAERTS L., STAMMINGER M., THEOBALT C.: Real-time expression transfer for facial reenactment. *ACM Trans. Graph. 34*, 6 (2015), 183:1–183:14. 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 16, 18, 27, 28

[TZS*16a] THIES J., ZOLLHÖFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2Face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (2016), CVPR '16, IEEE Computer Society. 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 26

[TZS*16b] THIES J., ZOLLHÖFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. *ArXiv, non-peer-reviewed prepublication by the authors abs/1610.03151* (2016). URL: http://arxiv.org/abs/1610.03151. 3, 6, 8, 9, 10, 16, 17

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. *ACM Trans. Graph. 24*, 3 (2005), 426–433. 3, 5, 9, 15, 26

[VIC] VICON: Vicon motion systems ltd. https://www.vicon.com. 3

[VJ04] VIOLA P., JONES M. J.: Robust real-time face detection. *Int. J. Comput. Vision 57*, 2 (2004), 137–154. 9

[VL08] VALLET B., LÃL'VY B.: Spectral geometry processing with manifold harmonics. *Computer Graphics Forum (Proceedings Eurographics)* (2008). 7

[VWB*12] VALGAERTS L., WU C., BRUHN A., SEIDEL H.-P., THEOBALT C.: Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph. 31*, 6 (2012), 187:1–187:11. 27, 28

[WBG*16] WU C., BRADLEY D., GARRIDO P., ZOLLHÖFER M., THEOBALT C., GROSS M., BEELER T.: Model-based teeth reconstruction. *ACM Trans. Graph. 35*, 6 (2016), 220:1–220:13. 13

[WBGB16] WU C., BRADLEY D., GROSS M., BEELER T.: An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph. 35*, 4 (2016), 115:1–115:12. 3, 4, 5, 6, 8, 9, 26

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Realtime performance-based facial animation. *ACM Trans. Graph. 30*, 4 (2011), 77:1–77:10. 3, 4, 5, 9, 10, 12, 15, 26

[WBM*16] WOOD E., BALTRUSAITIS T., MORENCY L. P., ROBINSON P., BULLING A.: A 3d morphable eye region model for gaze estimation. In *ECCV* (2016). 13

[WGP*10] WILSON C. A., GHOSH A., PEERS P., CHIANG J.-Y., BUSCH J., DEBEVEC P.: Temporal upsampling of performance geometry using photometric alignment. *ACM Trans. Graph. 29*, 2 (2010), 17:1–17:11. 3

[WHL*04] WANG Y., HUANG X., LEE C.-S., ZHANG S., LI Z., SAMARAS D., METAXAS D., ELGAMMAL A., HUANG P.: High resolution acquisition, learning and transfer of dynamic 3-D facial expressions. *Computer Graphics Forum 23*, 3 (2004), 677–686. 3

[Wil90] WILLIAMS L.: Performance-driven facial animation. *SIGGRAPH Comput. Graph. 24*, 4 (1990), 235–242. 1

[WLVGP09] WEISE T., LI H., VAN GOOL L., PAULY M.: Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2009), SCA '09, ACM, pp. 7–16. 7, 15

[WSVT13] WU C., STOLL C., VALGAERTS L., THEOBALT C.: Onset performance capture of multiple actors with a stereo camera. *ACM Trans. Graph. 32*, 6 (2013), 161:1–161:11. 28

[WSXC16] WANG C., SHI F., XIA S., CHAI J.: Realtime 3d eye gaze animation using a single rgb camera. *ACM Trans. Graph. 35*, 4 (2016), 118:1–118:14. 3, 8, 10, 12, 13

[WWMT11] WU C., WILBURN B., MATSUSHITA Y., THEOBALT C.: High-quality shape from multi-view stereo and shading under general illumination. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2011), CVPR '11, IEEE Computer Society, pp. 969–976. 7, 27

[WXLY17] WEN Q., XU F., LU M., YONG J.: Real-time 3d eyelids tracking from semantic edges. *ACM Trans. Graph. 36*, 6 (2017), 193:1–193:11. 13

[WZN*14] WU C., ZOLLHÖFER M., NIESSNER M., STAMMINGER M., IZADI S., THEOBALT C.: Real-time shading-based refinement for consumer depth cameras. *ACM Trans. Graph. 33*, 6 (2014), 200:1–200:10. 7

[XBMK04] XIAO J., BAKER S., MATTHEWS I., KANADE T.: Real-time combined 2d+3d active appearance models. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2004), CVPR'04, IEEE Computer Society, pp. 535–542. 9, 26

[XT13] XIONG X., TORRE F. D. L.: Supervised descent method and its applications to face alignment. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2013), CVPR '13, IEEE Computer Society, pp. 532–539. 26

[XWW*14] XU Z., WU H.-T., WANG L., ZHENG C., TONG X., QI Y.: Dynamic hair capture using spacetime optimization. *ACM Trans. Graph. 33*, 6 (2014), 224:1–224:11. 14

[ZCPR03] ZHAO W., CHELLAPPA R., PHILLIPS P. J., ROSENFELD A.: Face recognition: A literature survey. *ACM Comput. Surv. 35*, 4 (2003), 399–458. 2

[ZCW*17] ZHANG M., CHAI M., WU H., YANG H., ZHOU K.: A data-driven approach to four-view image-based hair modeling. *ACM Trans. Graph. 36*, 4 (2017), 156:1–156:11. 14

[Zha00] ZHANG Z.: A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 11 (2000), 1330–1334. 26

[ZR12] ZHU X., RAMANAN D.: Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2012), CVPR '12, IEEE Computer Society, pp. 2879–2886. 26

[ZSCS04] ZHANG L., SNAVELY N., CURLESS B., SEITZ S. M.: Spacetime faces: High resolution capture for modeling and animation. *ACM Trans. Graph. 23*, 3 (2004), 548–558. 3

[ZTCS99] ZHANG R., TSAI P.-S., CRYER J. E., SHAH M.: Shape-from-shading: a survey. *Trans. PAMI 21*, 8 (1999), 690–706. 7

[ZZZ15] ZAFEIRIOU S., ZHANG C., ZHANG Z.: A survey on face detection in the wild. *Comput. Vis. Image Underst. 138*, C (2015), 1–24. 9

Appendix A: Camera Models

In the literature, we can mainly identify three sorts of projection models: Orthographic, weak perspective, and full perspective.

Orthographic Projection

The orthographic projection model is a form of parallel projection, where optical rays are orthogonal to the camera plane (see Fig. 26 (a)). Here, every plane of the observed object appears in affine transformation on the viewing surface. This projection model is implicitly used by some 2D landmark tracking algorithms [CET01, MB04, XBMK04, XT13, ZR12] that directly operate on the 2D camera plane, and only by a few dense reconstruction algorithms [BY95, GRA13, EBDP96] due to its inability to handle input filmed with small focal length, which is a typical setting if webcams or RGB-D cameras are used.

In homogeneous coordinates, the orthogonal projection operator $\Pi_O(\cdot)$ is defined as a linear transformation:

$$\Pi_O(\cdot) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} . \tag{20}$$

The matrix of intrinsic parameters is simply defined as the identity matrix, i. e., $\mathbf{K} = \mathbf{I}$.

Weak Perspective Projection

The weak perspective projection model is commonly used by recent 2D landmark tracking algorithms [AZCP13, CC06, SLC11a] and has also been employed by some dense tracking approaches [BB01, LRF93, SWTC14, VBPP05] as it offers a good trade-off between simplicity and accuracy.

In this case, optical rays are orthogonal to the camera plane up to a scaling factor (see Fig. 26 (a)). The weak perspective projection is realized through a linear operator $\Pi_W(\cdot) = \rho\Pi_O(\cdot)$, where $\rho = 1/d$ is the scaling factor that accounts for global changes in depth d. Thus, 3D objects are approximated as planar surfaces that appear bigger or smaller in the 2D projection depending on their distance to the camera.

To represent a pixel $\mathbf{p}$ in homogeneous image coordinates, the matrix of intrinsic parameters $\mathbf{K}$ is defined as follows:

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & \mathbf{c}_x \\ 0 & 1 & \mathbf{c}_y \\ 0 & 0 & 1 \end{pmatrix} , \tag{21}$$

where $\mathbf{c} = [\mathbf{c}_x, \mathbf{c}_y]^\top$ is the principal point and represents the intersection between the optical axis and the image plane of the camera.

Full Perspective Projection

The full perspective model is more sophisticated and better reflects the working of a camera and the human eye. It is commonly used by dense tracking algorithms [BBPV03, BWP13, CHZ14, DM96, LYYB13, PSS99, SLL16, TZS*16a, WBLP11] and high-quality facial performance capture methods [CBZB15, GZC*16, IBP15, SSS14, WBGB16]. Here, optical rays converge at the camera and the projective geometry is mainly determined by the focal length f and the principal point $\mathbf{c} = [\mathbf{c}_x, \mathbf{c}_y]^\top$ (see Fig. 26 (b)). For the sake of simplicity, let us first assume that the principal point lies at $\mathbf{c} = [0,0]^\top$. By using similarity of triangles, we can associate a 3D point $\hat{\mathbf{v}}$ with a pixel $\hat{\mathbf{p}}$ in the sensor optics as follows:

$$\frac{\hat{\mathbf{p}}_x}{\hat{\mathbf{v}}_x} = \frac{\hat{\mathbf{p}}_y}{\hat{\mathbf{v}}_y} = \frac{f}{\hat{\mathbf{v}}_z} . \tag{22}$$

In this camera model, $\hat{\mathbf{v}}$ undergoes a non-linear perspective projection up to a factor given by the focal length f:

$$\hat{\mathbf{p}}_x = f\frac{\hat{\mathbf{v}}_x}{\hat{\mathbf{v}}_z}, \quad \hat{\mathbf{p}}_y = f\frac{\hat{\mathbf{v}}_y}{\hat{\mathbf{v}}_z} . \tag{23}$$

If we represent this transformation in homogeneous coordinates, the point $\hat{\mathbf{v}}$ is first projected using the non-linear operator $\Pi_F(\hat{\mathbf{v}}) = [\hat{\mathbf{v}}_x/\hat{\mathbf{v}}_z, \hat{\mathbf{v}}_y/\hat{\mathbf{v}}_z, 1]^\top$. Then, to properly represent a 2D point $\mathbf{p}$ in the camera plane under an arbitrary optical center, the matrix of intrinsic parameters $\mathbf{K}$ is defined as follows:

$$\mathbf{K} = \begin{pmatrix} f & 0 & \mathbf{c}_x \\ 0 & f & \mathbf{c}_y \\ 0 & 0 & 1 \end{pmatrix} . \tag{24}$$

The focal length f can be calibrated beforehand [BK13, Zha00] or roughly estimated based on the detected 2D landmarks [CWLZ13, CHZ14, GZC*16]. All three models do not represent optical distortions, such as spherical distortion, induced by real camera lenses. Therefore, such distortions need to be estimated during camera calibration and are then compensated by warping the images.

Appendix B: Light Transport

According to the well-known rendering equation, the light L radiating from a surface point $\mathbf{v}$ in the viewing direction ω_{out} can be expressed as

$$L(\mathbf{v}, \omega_{out}) = \int_\Omega L_{in}(\mathbf{v}, \omega_{in}) \cdot f_r(\mathbf{v}, \omega_{in}, \omega_{out}) \cdot \langle \mathbf{n}_\mathbf{v}, \omega_{in} \rangle \, d\omega_{in} , \tag{25}$$

where $\mathbf{n}_\mathbf{v}$ is the surface normal at point $\mathbf{v}$, ω_{in} is the direction of the incident light, and f_r is the BRDF (Bidirectional Reflectance Distribution Function), which describes how much light incident from direction ω_{in} is scattered to outgoing directions ω_{out}.

For purely diffuse surfaces, f_r is constant, and Eqn. 25 boils down to:

$$L(\mathbf{v}, \omega_{out}) = \rho(\mathbf{v}) \int_\Omega L_{in}(\mathbf{v}, \omega_{in}) \cdot \langle \mathbf{n}_\mathbf{v}, \omega_{in} \rangle \, d\omega_{in} = \rho(\mathbf{v}) D(\mathbf{v}) , \tag{26}$$

where $\rho(\mathbf{v})$ is the surface albedo and D the irradiance, i.e., the integrated incident light. This simplified lighting model is also referred to as *Lambertian reflection*.

To describe specular, i.e., non-diffuse, reflection, a large variety of BRDFs has been proposed. Common choices are the Phong and Blinn-Phong BRDFs [GGG*16], which are easy and efficient to compute, but are not physically based as they lack, for example, energy conservation. More physically plausible alternatives are microfacet models, such as the Torrance-Sparrow model. In practice, the specular component is often omitted to reduce the problem complexity, especially if real-time performance is required.

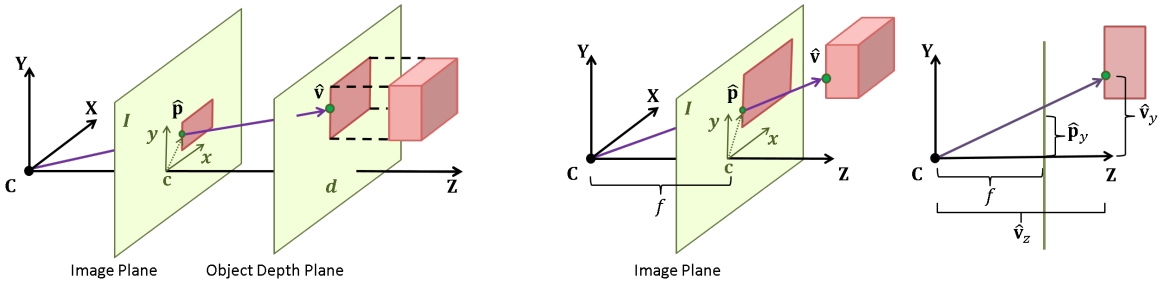

Figure 26: *Camera models: (a) Orthographic/weak perspective projection. The 3D object's geometry is flattened in depth (orthographic projection). The resulting geometry is rescaled based on its distance to the camera (weak perspective projection). (b) Full perspective projection. A 3D point* $\hat{\mathbf{v}}$ *is projected onto the image plane at position* $\hat{\mathbf{p}}$ *via a non-linear operator. Here, the image coordinate system is spanned by the vectors x and y. The intrinsic properties are given by the principal point* $\mathbf{c} = [\mathbf{c}_x, \mathbf{c}_y]^\top$ *(a,b), and also by the focal length f (b).*

Illumination Models

To simulate real world incident illumination on the face, a virtual illumination model is required. For tractability of the problem, many different assumptions and simplifications are made. A simple illumination model uses a single directional light source for direct lighting, and models indirect lighting with a constant ambient illumination term. More complex illumination conditions are described as a spherical function, stored explicitly in a texture (environment map) or by using appropriate basis functions (spherical harmonics).

Environment Maps

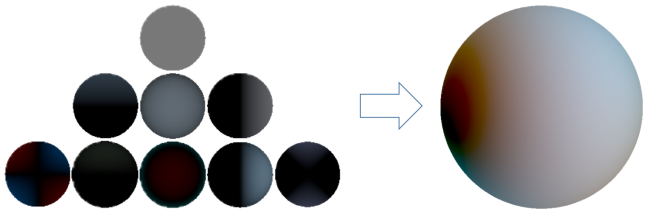

Figure 27: *Example of SH illumination on a sphere. Left: first three bands of the spherical harmonics; Right: composition of the bands.*

An environment map stores the incident light L_{in} in a texture image [AFB*13]. This technique usually assumes distant lighting and no self-shadowing, i. e., the incident light from direction ω_{in} is approximately the same at every point [KRP*15] and thus only depends on ω:

$$L_{in}(\mathbf{v}, \omega) = L_{in}(\omega) \quad . \tag{27}$$

Environment maps can be represented in different ways, e. g., by a cube map that remaps the sphere of incident directions onto the faces of a cube. In this case, six textures are stored, each containing n^2 pixels. To get $L_{in}(\omega_{in})$, the textures must be sampled w. r. t. the direction ω_{in}. The higher the resolution of the textures, the better the approximation of the environment. However, this representation has $6 \cdot n^2$ unknowns per color channel, which is rather high and makes the inverse rendering problem computationally infeasible. Another more compact, yet powerful, way to approximate environment maps is to represent them in a spherical function basis, in particular the so-called spherical harmonics basis.

Spherical Harmonics

Spherical harmonics (SH) are orthogonal basis functions defined on polar coordinates and represent a specialization of the Legendre polynomials [Mue66]. An example of an illumination estimate represented using SH is shown in Fig. 27. SH are organized in bands and the real basis of spherical harmonics can be calculated by the following formula:

$$Y_l^m(\theta, \phi) = \begin{cases} \sqrt{2} K_l^m \cdot \cos(m \cdot \phi) \cdot P_l^m(\cos(\theta)) & \text{if } m > 0 \\ K_l^m \cdot P_l^m(\cos(\theta)) & \text{if } m = 0 \\ \sqrt{2} K_l^m \cdot \sin(-m \cdot \phi) \cdot P_l^{-m}(\cos(\theta)) & \text{if } m < 0 \end{cases} , \tag{28}$$

where l is the index of the band ($l \in \mathbb{N}_0$) and m is the index within a band ($m \in [-l, l]$). Here, the K_l^m denote the normalization constants and P_l^m are the associated Legendre polynomials. Tab. 2 shows the first four bands of the SH basis functions in Cartesian coordinates.

A function $g(\omega)$ that is defined on a sphere, e. g., a spherical environment map, can then be approximated using these basis functions, as follows:

$$g(\omega) \approx \sum_{l=0}^{b-1} \sum_{m=-l}^{l} g_l^m Y_l^m(\omega) \quad . \tag{29}$$

Here, b is the number of used bands and g_l^m are the coefficients of the corresponding basis functions. For the sake of simplicity, let us assume w. l. o. g. a linear order of the spherical harmonics instead of the band organization, i. e., $Y_1 = Y_0^0, Y_2 = Y_1^{-1}, Y_3 = Y_1^0, ..., Y_{b^2} = Y_{b-1}^{b-1}$. This means that b^2 coefficients are required to describe the environment map using the first b SH bands.

In our context, it is mostly assumed that faces are purely diffuse, i. e., Lambertian, objects [GVWT13, GZC*16, SSS14, TZN*15, VWB*12, WWMT11]. As shown in Eqn. 26, this assumption yields a Rendering Equation that only depends on the position $\mathbf{v}$ and not

Band (l)	Index within a band (m)						
	-3	-2	-1	0	1	2	3
0				$\frac{1}{2\sqrt{\pi}}$			
1			$\frac{\sqrt{3}}{2\sqrt{\pi}}y$	$\frac{\sqrt{3}}{2\sqrt{\pi}}z$	$\frac{\sqrt{3}}{2\sqrt{\pi}}x$		
2		$\frac{\sqrt{15}}{4\sqrt{\pi}}(x^2-y^2)$	$\frac{\sqrt{15}}{2\sqrt{\pi}}xz$	$\frac{\sqrt{5}}{4\sqrt{\pi}}(3z^2-1)$	$\frac{\sqrt{15}}{2\sqrt{\pi}}yz$	$\frac{\sqrt{15}}{2\sqrt{\pi}}xy$	
3	$\frac{\sqrt{35}}{\sqrt{32\pi}}(x^2-3y^2)x$	$\frac{\sqrt{105}}{4\sqrt{\pi}}(x^2-y^2)z$	$\frac{\sqrt{21}}{\sqrt{32\pi}}(4z^2-x^2-y^2)x$	$\frac{\sqrt{7}}{4\sqrt{\pi}}(2z^2-3x^2-3y^2)z$	$\frac{\sqrt{21}}{\sqrt{32\pi}}(4z^2-x^2-y^2)y$	$\frac{\sqrt{105}}{2\sqrt{\pi}}(xyz)$	$\frac{\sqrt{35}}{\sqrt{32\pi}}(3x^2-y^2)y$

Table 2: *The first four bands of the spherical harmonics basis functions in Cartesian coordinates.*

on the viewing direction ω:

$$L(\mathbf{v},\omega_{out}) = \rho(\mathbf{v}) \cdot D(\mathbf{v}) \quad . \tag{30}$$

Computing $D(\mathbf{v})$ still requires an integration over the entire hemisphere Ω, to sum-up the incident light L_{in}. Even though highly optimized sampling strategies have been examined to perform this integration, it may still be inefficient and inaccurate. Instead, state-of-the-art methods first convert the environment map (storing $L_{in}(\omega)$) to another map storing the Lambertian reflection $D(\mathbf{v})$, e. g., [VWB*12, WSVT13, TZN*15]. Using the assumption of distant lighting (cp. Eqn. 27), $D(\mathbf{v})$ is approximated by a spherical function $\tilde{D}(\mathbf{n_v})$ that only depends on the surface normal $\mathbf{n_v}$. $\tilde{D}(\mathbf{n_v})$ can be computed from $L_{in}(\omega)$ by the convolution with a cosine-function [RH01a, RH01b]. Due to this convolution, $\tilde{D}$ is much smoother than L_{in} and can be well approximated using SH:

$$\tilde{D}(\mathbf{n_v}) \approx \sum_{i=1}^{b^2} l_i Y_i(\mathbf{n_v}) \quad . \tag{31}$$

Following [RH01a, RH01b], $\tilde{D}$ is so smooth that three SH bands $b = 3$ are sufficient to achieve an average error below 1%, independent of the environment map. This results in only $m_l = b^2 = 9$ variables per color channel. An important property of the SH lighting model is that it is differentiable and fast to evaluate.