Facial Performance Capture from Visual Input and EMG Signals

Jianwen Lou

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth.

September 2020

Abstract

Facial performance capture (or face capture), a process of reconstructing, tracking and analysing the deformable geometry and appearance of the human face from visual input (e.g. RGB or RGB-D images), is a long-standing research topic in the fields of computer graphics and vision. Over the past two decades, the field of face capture has witnessed rapid progress, which has pushed the capture method's accuracy, speed and ease of use to a new level, and benefited a wide range of applications such as personalized facial avatar generation, face identification and facial animation. Nevertheless, it remains an open problem of improving the method's capturing robustness while keeping the method compute and data-efficient. Moreover, the emerging virtual reality (VR) technologies for immersive interactions have posed new challenges to facial performance capture. The VR head-mounted display (HMD) occludes a large portion of the user's face, which makes conventional vision-based face capture methods less effective.

Targeting at solving the aforementioned problems, this thesis first develops two novel face capture approaches for detecting sparse 2D facial landmarks and tracking dense 3D facial geometry respectively from a monocular RGB camera. Both approaches have been thoroughly evaluated on benchmark face image and video datasets. In comparison with the previous methods, they showcase improved capturing performance at very low data and computational cost. The proposed approaches have further been implemented into mobile and desktop facial tracking interfaces and validated on live video streams.

For capturing the VR HMD user's facial expression with high-fidelity, the thesis proposes to combine a classic monocular 3D face reconstruction algorithm with a pioneering facial biosensing technique – Faceteq, which uses advanced electromyographic (EMG) sensors to capture facial muscle activities. This extends the facial performance capture from the traditional visual scene to the novel VR

context, thereby providing a practical solution to achieve face-to-face communication with compelling facial expressions in virtual environment.

Besides developing robust facial performance capture approaches, this thesis explores a new direction for applying those approaches to solve real-world problems. Specifically, it identifies the problem of automated facial nerve function assessment from visual face capture for facial palsy management. By systematically reviewing the principal studies on related topics, the thesis points out the challenges in the field and indicates promising directions for future work. What's more, it proposes a promising pathway to apply the face capture methods proposed in previous chapters onto automated facial nerve function assessment. To the best of my knowledge, this is the first review of its kind to be reported so far. Due to the interdisciplinary nature of the review, it can benefit multiple areas, including visual face capture, clinical facial palsy diagnosis and facial bioengineering.

Contents

Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

List of Figures

List of Tables

List of Acronyms

3DMM	3D Morphable Model
AU	Action Unit
BFM	Basel Face Model
CNN	Convolutional Neural Network
EEG	Electroencephalography
EMG	Electromyography
ESFP	Emotionally Salient Facial Part
ESR	Explicit Shape Regression
FACS	Facial Action Coding System
GAN	Generative Adversarial Network
GoMBF	Globally-optimized Modular Boosted Ferns
GoMBF-Cascade	Cascade of Globally-optimized Modular Boosted Ferns
GSDM	Global Supervised Descent Method
HDR	High Dynamic Range
HMD	Head-mounted Display
MS-SDM	Multi-subspace Supervised Descent Method
nsINST	Non-semantic Instrument
PCA	Principal Component Analysis
RMS	Root Mean Square
RMSE	Root Mean Square Error
SDM	Supervised Descent Method
sINST	Semantic Instrument
VR	Virtual Reality

Acknowledgements

Foremost, I would like to greatly thank my supervisor Prof. Hui Yu for providing this wonderful PhD opportunity and giving me constant support over the past four years. I also would like to stress our great and successful cooperation with Emteq. The professionalism of the colleagues in Emteq has deeply impressed me. In particular, I learned a lot from Charles Nduka who is always with full enthusiasm to new knowledge and has many insightful ideas.

To work and to party with my colleagues were marvellous. I would like to thank Qiongdan Cao, Yifan Xia, Hao Fan, Jianyuan Sun, Yiming Wang, Weihong Gao, Martin Kearl and Shu Zhang. We have many good moments together in Portsmouth.

Last but not least, I dedicate this thesis to my wife Xiaoxu Cai and my parents with love. They are staying by my side all the way through this journey.

Dissemination

Lou, J., Cai, X., Dong, J., & Yu, H. (2020). Real-time 3D facial tracking via cascaded compositional learning. *IEEE Transactions on Image Processing*, under review. (I proposed the facial motion regression method, did the experimental validation and wrote up the paper.)

Lou, J., Wang, Y., Nduka, C., Hamedi, M., Mavridou, I., Wang, F. Y., & Yu, H. (2019). Realistic facial expression reconstruction for VR HMD users. *IEEE Transactions on Multimedia*, *22*(3), 730–743. (I designed and implemented most parts of the reconstruction system, including 3D face embodiment generation, FACS AUs prediction from EMG signals and emotion recognition from AUs. I also conducted the experimental validation and wrote up the paper.)

Lou, J., Yu, H., & Wang, F. Y. (2019). A review on automated facial nerve function assessment from visual face capture. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(2), 488–497. (I conducted the investigation of the target problem and wrote up the paper.)

Lou, J., Cai, X., Wang, Y., Yu, H., & Canavan, S. (2019). Multi-subspace supervised descent method for robust face alignment. *Multimedia Tools and Applications*, *78*(24), 35455–35469. (I proposed the multi-subspace supervised descent method, conducted the experimental validation and wrote up the paper.)

Xia, Y., **Lou, J.**, Dong, J., Qi, L., Li, G., & Yu, H. (2020). Hybrid regression and isophote curvature for accurate eye center localization. *Multimedia Tools and Applications*, 79(1), 805-824.

Cai, X., Yu, H., **Lou, J.**, Zhang, X., Li, G., & Dong, J. (2020). 3D facial geometry recovery from a single depth view with attention-guided GAN. *IEEE Transactions on Cognitive and Developmental Systems*, under review.

Chapter 1
Introduction

1.1 Background

The face occupies a central position in communicating social information such as identity, emotion and intent between humans. This has inspired a long-standing research topic in computer graphics and vision communities, which focuses on developing technical solutions for reconstructing, tracking and analysing the deformable geometry of a human face as well as its texture from visual input (e.g. RGB or RGB-D images), a process typically referred to as facial performance capture (or face capture) (Zollhöfer et al., 2018). The application domain of facial performance capture is vast, ranging from facial recognition for intelligent human-machine interface, personalized avatars for facial virtual reality in entertainment and social media, facial modification (e.g. touch-up, completion and reenactment) for visual effects in high-end productions such as movies and computer games, all the way to facial biometrics for medicine and healthcare.

A fundamental step in facial performance capture is to detect sparse 2D facial landmarks (see Fig. 1.1) given an image. This process is normally called as face alignment (Xiong & De la Torre, 2013). The detected 2D landmarks explicitly outline the 2D facial shape, hence can facilitate tasks such as facial expression recognition in near-frontal poses and provide good priors to constrain the ill-posed monocular 3D face reconstruction problem. However, the 2D facial representation is limited in depicting the out-of-plane rotation and the unseen facial texture when there exists a head pose. Recovering the dense models (Thies, Zollhofer, et al., 2016) of 3D facial geometry and appearance (see Fig. 1.1) from visual data is thus the key technology required in the vast majority of applications. On the sensor side, there are various setups can be employed based on the target application and the available resources, such as the multi-camera setup with controlled lights, the

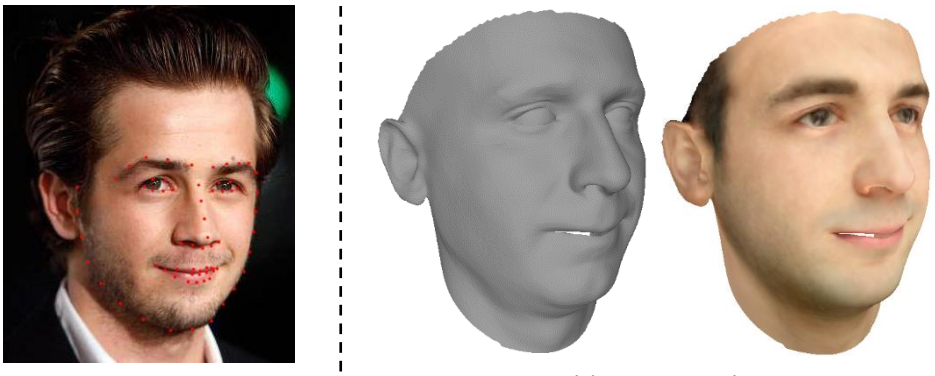
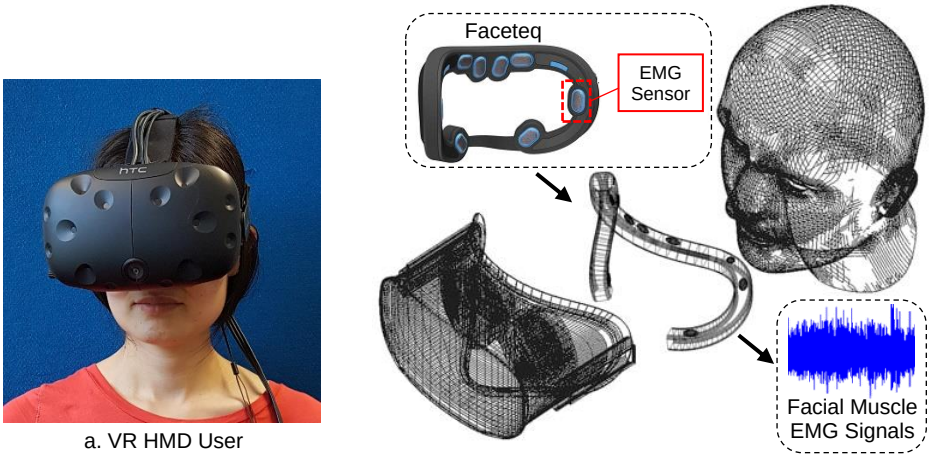


Figure 1.1: 2D facial landmarks and 3D facial geometry and appearance.

depth camera or a single commodity RGB camera. Fuelled by the increasing demand for consumer face technologies and even the professional face related tasks in content creation pipelines which are still highly manual and laborious at this stage, face capture with a ubiquitous monocular RGB camera has become the major research interest in recent years (Zollhöfer et al., 2018). The thesis is motivated by this important tendency, centred on developing robust 2D and 3D facial performance capture solutions with the most lightweight capture setup – a single monocular RGB camera.

1.2 Problems and Challenges

In the past two decades, the field of facial performance capture from monocular RGB input has witnessed remarkable progress with a series of novel and powerful methods proposed (Zollhöfer et al., 2018). **However, it remains challenging to attain a promising capturing result with a small training set (e.g. the training image amount is less than 5K), or when trying to achieve low computational complexity for making the capture method more applicable to consumer applications (e.g. mobile applications)**. Existing methods that cope with the challenging cases such as big head pose and poor illumination, normally rely on either a large-scale training set (Y. Guo, Zhang, Cai, Jiang, et al., 2018) or a highly-



a. VR HMD User

b. Faceteq for Detecting Facial Muscle Activity

Figure 1.2: VR HMD user and Faceteq (Mavridou et al., 2017).

complicated photo-geometric fitting process (Thies, Zollhofer, et al., 2016). They hence can hardly be applied to scenarios where the labelled training data is difficult to acquire or the target platform's computing resource is limited. On the contrary, the methods (C. Cao et al., 2014; Huber et al., 2016) that pursue the training or run-time efficiency always produce compromised reconstruction and tracking results. **What's more, the traditional way of harvesting labelled training data is expensive and laborious, which severely hinders the development of capture approaches.** In 3D face capture, this is particularly problematic as the 3D label is normally obtained by multi-view stereo (Laine et al., 2017), photometric (Y. Guo, Zhang, Cai, Jiang, et al., 2018) or 2D landmark-based (C. Cao et al., 2014) reconstruction which requires either complicated and expensive multi-camera setups or laborious manual annotations. Thus, a more economic and efficient data collection way such as synthesizing training imagery is highly needed.

In virtual reality (VR), which provides distinctive immersive interaction and is turning into a next-generation communication platform, the vision-based capture approaches will become less effective as the user's face is severely occluded by the VR head-mounted display (HMD, see Fig. 1.2a). **This makes the face capture for VR HMD users very challenging and hinders the users to communicate**

face-to-face with compelling facial expressions and emotions in virtual environments. Despite its importance, the problem has drawn little attention in research communities, resulting in no valid solutions that can be applied to general VR settings have been proposed yet.

Along with the necessity of developing novel technical solutions for improving facial performance capture in a broader range of scenarios, **exploring new avenues to apply these technological outcomes is also crucial but always being neglected**. There are plenty of applications on smart human-machine interface and the content creation for entertainment purpose such as movies and gaming, while only a small portion of applications are designed for facial biometrics on medicine and heath, especially the physical health aspect.

1.3 Contributions and Outline

With the primary goal of tackling the aforementioned problems and challenges, the rest of this thesis is unfolded into five chapters and mainly makes the following contributions:

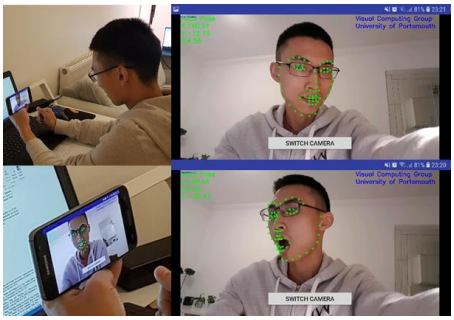
- In Chapter 2 and Chapter 3, it develops two novel visual tracking algorithms for capturing 2D (Lou, Cai, et al., 2019) and 3D (Lou et al., 2020) facial motion respectively from the monocular RGB input. The developed algorithms are demonstrated to be not only resistant to adverse tracking conditions, but also compute and data-efficient. The two algorithms both adopt the framework of cascaded regression but with different emphasis: the former algorithm concentrates on regressing unimodal 2D facial landmarks from image features, while the latter one focuses on regressing multi-modal 3D facial motion parameters. The 2D facial landmarks output from the first algorithm serve as the essential prior information when solving the problems of 3D facial tracking and reconstruction in Chapter 3 and Chapter 4.
- For an improved data collection way, Chapter 3 also deeply investigates the effect of synthesized facial images in training the 3D facial motion regressor.

Synthesizing training imagery is purely automatic and highly efficient, hence providing a novel and valuable direction for collecting the training data.

- Chapter 4 extends the facial performance capture from the traditional visual scene to the VR context that affords unique immersive interaction via the HMD. It proposes to combine a monocular 3D face reconstruction algorithm with, a pioneering biosensing technique – Faceteq^{TM} (see Fig. 1.2b) (Mavridou et al., 2017) which employs electromyography (EMG) to detect facial muscle activities and can be seamlessly integrated into mainstream HMDs, to enable high-fidelity facial expression capture for VR HMD users (Lou, Wang, et al., 2019). The 3D face reconstruction algorithm applied in this chapter is an enhanced version as that used in Chapter 3. Chapter 3's reconstruction algorithm utilizes only facial landmark information, while Chapter 4's algorithm further incorporates image colour information into the optimization process for recovering high-fidelity facial texture.
- Last but not least, Chapter 5 explores a novel application avenue – automated facial nerve function assessment from visual face capture for the methods proposed in previous chapters. Specifically, it systematically reviews the most relevant and important studies in the area, indicates challenges and new directions on utilizing face capture outcomes for more efficient and objective facial palsy management (Lou, Yu, et al., 2019).

In the following, a more detailed abstract of each chapter is given:

Chapter 2 - Robust 2D Face Alignment with Multi-subspace Supervised Descent Method:



Supervised Descent Method (SDM) (Xiong & De la Torre, 2013) is a leading cascaded regression method for face alignment, which achieves the state-of-the-art performance and has a solid theoretical basis. However, SDM is prone to local optima and unable to handle

the optimization space of face alignment with conflicting descent maps. This limits SDM in unconstrained face alignment which deals with face images of various facial shapes and appearance. In this chapter, a novel two-step method called multi-subspace SDM (MS-SDM) is proposed to equip SDM with a stronger capability of coping with the unconstrained faces. MS-SDM first partitions the original optimization space of face alignment into several subspaces by applying k-means on facial appearance features. The generated subspaces show a clear semantic link to the head pose. Then, it learns an independent feature-shape regression for each subspace via SDM. During testing, the face image will be assigned into the correct subspace with a robust Naive Bayes classifier and the corresponding shape regression will be called to detect landmarks. MS-SDM has been evaluated on benchmark face datasets and live video streams with a mobile facial tracking implementation. It shows improved landmark detection performance comparing with SDM and its variant GSDM.

Chapter 3 - Real-time 3D Facial Tracking via Cascaded Compositional Learning:

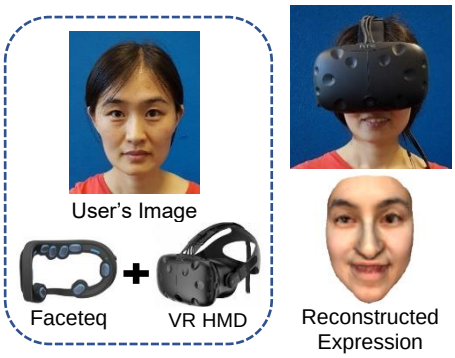
This chapter proposes to learn a cascade of globally-optimized modular boosted ferns (GoMBF) to solve multi-modal facial motion regression for real-time 3D facial tracking from a monocular RGB camera. GoMBF is a deep composition of multiple regression models with each is a boosted ferns initially trained to predict partial motion parameters of the same modality, and then concatenated together via a global optimization step to form a singular strong boosted ferns that can effectively handle the whole regression target. It can explicitly cope with the modality variety in output variables, while manifesting increased fitting power and a faster learning speed comparing against the conventional boosted ferns. By further cascading a sequence of GoMBFs (GoMBF-Cascade) to regress facial motion parameters, it outputs competitive tracking performance on a variety of in-the-wild videos

comparing to the state-of-the-art methods, which require much more training data or have higher computational complexity. It provides a robust and highly elegant solution to real-time 3D facial tracking using a small set of training data and hence makes it more practical in real-world applications.

Comparing with traditional manual collection, synthetic generation of training imagery provides a more economic and efficient data collection way. It offers a promising solution to mitigate the training data shortage in 3D facial tracking. However, it remains unclear to what extent the synthetic data has contributed to training the tracking model. To solve this problem, this chapter then deeply investigates the effect of synthesized facial images on training GoMBF-Cascade for 3D facial tracking. It applies three types synthetic images with various naturalness levels for training, and compares the performance of the tracking models trained on real data, on synthetic data and on a mixture of data. The experimental results indicate that, i) the model trained purely on synthetic facial imageries can hardly generalize well to unconstrained real-world data, ii) involving synthetic faces into training benefits tracking in some certain scenarios but degrades the tracking model's generalization ability. These two insights could benefit a range of non-deep learning facial image analysis tasks where the labelled real data is difficult to acquire.

Chapter 4 - Realistic 3D Facial Expression Reconstruction for VR HMD Users:

This chapter develops a system for sensing and reconstructing facial expressions of the VR HMD user. The HMD occludes a large portion of the user's face, which makes most existing facial performance capture techniques intractable. To tackle this problem, a novel hardware solution with EMG sensors being attached to the headset frame is applied to track facial muscle movements. For realistic facial expression recovery, the developed system



first reconstructs the user's 3D face from a single image and generates the personalized blendshapes associated with seven facial action units (AUs) on the most emotionally salient facial parts (ESFPs). It then utilizes pre-processed EMG signals for measuring activations of AU-coded facial expressions to drive pre-built personalized blendshapes. Since facial expressions appear as important nonverbal cues of the subject's internal emotional states, the system further investigates the relationship between six basic emotions - anger, disgust, fear, happiness, sadness and surprise, and the detected AUs using a fern classifier. Experiments show the proposed system can accurately sense and reconstruct high-fidelity common facial expressions while providing useful information regarding the emotional state of the HMD user.

Chapter 5 - A Review on Automated Facial Nerve Function Assessment from Visual Face Capture: Assessing facial nerve function from visible facial signs such as resting asymmetry and symmetry of voluntary movement is an important means in clinical practice. By using image processing, computer vision and machine learning techniques, replacing the clinician with a machine to do assessment from ubiquitous visual face capture is progressing more closely to reality. This approach can do assessment in a purely automated manner, hence opens a promising direction for future development in this field. Many studies gathered around this interesting topic with a variety of solutions proposed in recent years. However, to date, none of these solutions have gained a widespread clinical use. This chapter provides a comprehensive review of the most relevant and representative studies in automated facial nerve function assessment from visual face capture. It deeply discusses the challenges and directions in developing the assessment method. Specially, it identifies the significance and potential of monocular face capture approaches in achieving fully automated, objective and accurate facial nerve function assessment. This introduces a promising avenue to apply and improve the face capture approaches proposed in previous chapters. At the end of the chapter, a pathway to implement such an application is proposed. To the best of my knowledge, this is the first study of its kind to be reported so far.

It can benefit multiple groups of people, ranging from visual face capture researchers to clinical practitioners.

Chapter 6 – Summary and Outlook: This chapter summarises the thesis with an in-depth discussion on its contributions and the future work.

Chapter 2

Robust 2D Face Alignment with Multi-subspace Supervised Descent Method

Foreword

This chapter develops a novel 2D face alignment method for detecting facial landmarks given an unconstrained face image. Face alignment is a fundamental task in facial performance capture. Its outcomes provide critical prior information to the challenging monocular 3D facial tracking and reconstruction problems which will be studied in the subsequent chapters of this thesis. The proposed method is based on the classic Supervised Descent Method which is a representative cascaded regression method.

The chapter is based on a published journal paper:

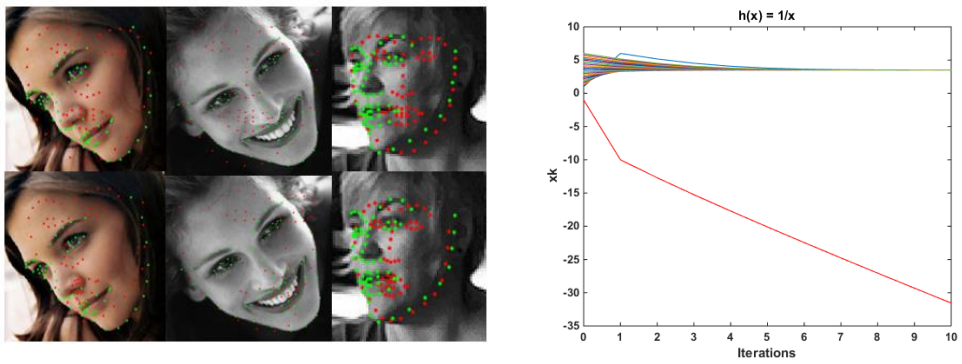
- **Lou, J.**, Cai, X., Wang, Y., Yu, H., & Canavan, S. (2019). Multi-subspace supervised descent method for robust face alignment. *Multimedia Tools and Applications*, 78(24), 35455-35469.

in which I proposed the multi-subspace supervised descent method, conducted the experimental validation of the proposed method, implemented the mobile 2D facial tracking application and wrote up the paper.

2.1 Introduction

Face alignment aims to locate facial landmarks which outline the 2D shapes of key facial parts such as the eyebrows, the eyes and the mouth in a face image. It is a fundamental step in many face-related tasks like 3D face modelling (C. Cao et al., 2014; Jiang et al., 2018), face frontalization (Y. Wang et al., 2016, 2017), and facial attributes prediction (Jian & Lam, 2015; Xia et al., 2018). The field of face alignment has witnessed rapid progress in recent years, especially after the advent of the cascaded regression method (X. Cao et al., 2014; Xiong & De la Torre, 2013; X. Zhu et al., 2016). Generally, the cascaded regression method learns a sequence of shape (a set of landmarks) increment from facial appearance features to progressively update the initial facial shape towards the ground-truth shape. A variety of method variants have been proposed in this important research stream, in which SDM (Xiong & De la Torre, 2013) is a very popular one. SDM achieves the state-of-the-art landmark detection accuracy while is extremely efficient in both the training and testing phases. What's more, it is theoretically sound to some extent based on a rigorous interpretation from the perspective of solving a nonlinear optimization problem with the Newton's method.

Despite its great success in face alignment, SDM has two main limitations: 1) It highly relies on the initialization and is prone to local optima. SDM is derived from the Newton's method which inherently finds a local minimum of the optimization problem. Therefore, if the initial landmarks are far away from the target landmarks, SDM will be easily trapped into a poor local optimum (see Fig. 2.1a). 2) It is unable to deal with the optimization problem that contains conflicting decent maps. To illustrate this problem, the feature extraction function applied in SDM is simply assumed to be $h(x) = x^{-1}$. Supposing that our aim is to seek from a range of initial $x(x_0)$ the optimal $x(x_* = 3.5)$ that makes $h(x) = 0.286$ (please note that x here has no specific meaning and x_* can be set as any other values as you wish). In SDM, a generic descent map r will be learned to update x_0 towards x_* by calling the following equation:



a. b.

Figure 2.1: The limitations of SDM. a. The failure cases of SDM due to poor initializations. Top row: initial landmarks, bottom row: updated landmarks after four iterations. Red points: predicted landmarks, green points: ground-truth landmarks. b. Initial points that have conflicting descent maps.

$$x_k = x_{k-1} - r\big(h(x_{k-1}) - h(x_*)\big) \tag{2.1}$$

As shown in Fig. 2.1b, with $r = -7$, all $x_0 \in [1\!:\!0.2\!:\!6]$ (0.2 is the interval) can be moved closer to x_*, while $x_0 < 0$ (e.g. $x_0 = -1$) will be moved farther away from x_*. In theory, only if the initial data points are close to each other and target at the same destination, SDM would be able to learn a generic descent map that applies to all the data points. This prerequisite however can hardly be met in practical face alignment tasks, where the facial shape and appearance (feature) vary a lot across different face images due to the variety of head poses, facial expressions and lighting conditions.

To tackle the limitations of SDM, this chapter develops an efficient and novel two-step method – multi-subspace SDM (MS-SDM, see Fig. 2.2). MS-SDM partitions the original optimization space of the face alignment problem into multiple subspaces with k-means. In each subspace, the faces exhibit similar shapes, thereby a valid generic descent map can be learned much more easily via SDM. During testing, a face image will be directed to the correct subspace with a pre-trained Naive Bayes classifier and the corresponding feature-shape regression will then be called to detect facial landmarks. The proposed MS-SDM has been

validated on challenging face datasets which cover a wide range of head poses, facial expressions and appearance. Experimental results show the superiority of MS-SDM over SDM and its variant – GSDM (Xiong & De la Torre, 2015).

2.2 Related Work

Existing face alignment methods can generally be divided into two main categories: generative approaches and discriminative approaches. Generative approaches, such as Active Appearance Models (Matthews & Baker, 2004) and Constrained Local Models (Cristinacce & Cootes, 2006), first construct the parametric facial shape and appearance models using Principal Component Analysis (PCA), then generate a model instance to fit with the face image via a nonlinear optimization step. This kind of method is normally compute-intensive and the expressive power of the parametric face model is supposed to be limited. Discriminative approaches directly learn a mapping (or regression) from image features to landmark locations. As a representative discriminative approach, cascaded regression (X. Cao et al., 2014; Xiong & De la Torre, 2013; S. Zhu et al., 2015, 2016) has dominated the field of face alignment in recent years due to its high computational efficiency and landmark detection accuracy.

2.2.1 Face Alignment with Cascaded Regression

Starting with a coarse initial shape, the cascaded regression method gradually refines the shape by estimating a shape increment stage-by-stage with a sequence of regression functions. Cao et al. (X. Cao et al., 2014) applied the boosted ferns to learn the features and the nonlinear feature-shape mappings simultaneously, which delivered promising landmark detection results. Xiong et al. (Xiong & De la Torre, 2013) instead employed a much simpler linear regression and the hand-crafted features to build the cascaded regression. Their method is named as Supervised Descent Method (SDM). Although built with a very simple setup, SDM achieved the state-of-the-art landmark detection performance. In more recent

studies (Saeed et al., 2018; X. Zhu et al., 2016), deep learning methods have become the major research interest. The strong feature learning ability and the end-to-end learning mode enable the deep learning method to output remarkable face alignment results even on the most challenging face datasets. However, deep learning methods normally require a large-scale training set and bear a high computational complexity, thus are difficult to be deployed to devices (e.g. mobile phones) with limited computing resources. For a more comprehensive review of the mainstream cascaded regression methods, readers are directed to surveys (Chrysos et al., 2018; N. Wang et al., 2018). This chapter chooses SDM as the building block of the proposed face alignment algorithm because of its extremely high efficiency in both the training data and the runtime prediction.

2.2.2 Face Alignment with SDM-based Approaches

SDM gains the state-of-the-art face alignment performance with very low data and computational cost. It has become a benchmark cascaded regression method and motivated a number of new approaches in face alignment. As discussed above, for learning a valid generic descent map, SDM requires the initial facial shape to lie close to the target shape and the feature-shape relationships of different face images should be close to each other. However, this prerequisite can hardly be satisfied in practical face alignment tasks, as the facial shape and appearance could differ significantly among different face images.

The issues above hinder SDM to cope with the unconstrained face alignment problem, which presumably occupies multiple optimization subspaces caused by the diversity of facial shape and appearance in face images, hence can hardly be addressed with a single optimization process as that employed in SDM. Xiong and De la Torre (Xiong & De la Torre, 2015) made a similar inference and proposed the global SDM – GSDM. GSDM first conducts domain partition in facial feature and shape PCA spaces, then applies the domain-specific SDM for more robust 2D facial tracking in a video. Nevertheless, GSDM requires the near ground-truth facial landmarks when selecting the domain, which prohibits it from dealing with

single image face alignment. The utilization of PCA also remains a big concern as it might cause an unestimated information loss. Zhang et al. (Y. Zhang et al., 2016) later improved the SDM by projecting both the facial shape and feature into a mutual sign-correlation subspace before performing the regression. Their method however has the same constraints as those found in GSDM. Alternatively, a few studies (Liu et al., 2015; X. Yu et al., 2016) resort to the multi-view solution - processing the face image with a view-specific shape regressor which was chosen based on the head pose estimation. This kind of method is capable of handling face images spanning a broad range of head poses. However, it neglects the non-rigid shape deformations caused by facial expressions and the rich appearance variations in face images.

2.3 Methodology

This section first recaps the SDM method and analyses its limitations. Then, it introduces the proposed MS-SDM.

2.3.1 Supervised Descent Method

Given a face image I and the initial facial shape (landmarks) $\mathbf{x}_0 \in \mathbb{R}^{2L}$ (L is the number of landmarks), face alignment can be framed as minimizing the following objective function over the shape increment $\Delta\mathbf{x}$:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|h(\mathbf{x}_0 + \Delta\mathbf{x}, I) - h(\mathbf{x}_*, I)\|_2^2 \tag{2.2}$$

where $h(\mathbf{x}, I) \in \mathbb{R}^{128L}$ represents the SIFT features (Lowe, 2004) extracted around the landmarks $\mathbf{x}$ from image I. $\mathbf{x}_*$ denotes the ground-truth landmark positions. Following the Newton's method with a second-order Taylor expansion, Eq. 2.2 can be rewritten as:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) \approx f(\mathbf{x}_0) + \mathbf{J}_f(\mathbf{x}_0)^{\mathrm{T}}\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^{\mathrm{T}}\mathbf{H}_f(\mathbf{x}_0)\Delta\mathbf{x} \tag{2.3}$$

where $\mathbf{J}_f(\mathbf{x}_0)$ and $\mathbf{H}_f(\mathbf{x}_0)$ are the Jacobian and Hessian matrices of f evaluated at $\mathbf{x}_0$. Differentiating Eq. 2.3 with respect to $\Delta\mathbf{x}$ and setting it to zero, we can obtain:

$$\begin{aligned}\Delta\mathbf{x} &= -\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_f(\mathbf{x}_0)\\ &= -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathrm{T}}(\mathbf{x}_0)\big(h(\mathbf{x}_0, I) - h(\mathbf{x}_*, I)\big)\\ &= -2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathrm{T}}(\mathbf{x}_0)h(\mathbf{x}_0, I) + 2\mathbf{H}_f(\mathbf{x}_0)^{-1}\mathbf{J}_h^{\mathrm{T}}(\mathbf{x}_0)h(\mathbf{x}_*, I)\end{aligned} \tag{2.4}$$

According to Eq. 2.4, calculating the shape increment $\Delta\mathbf{x}$ needs $h(\mathbf{x}, I)$ to be twice differentiable or the numerical approximations of the Jacobian and the Hessian can be calculated. However, these requirements are difficult to meet in practice: 1) SIFT or HOG features are non-differentiable image operators; 2) numerically estimating the Jacobian or the Hessian in Eq. 2.4 is computationally expensive. For example, calculating the inverse of Hessian matrix is with $\mathrm{O}(p^3)$ time complexity and $\mathrm{O}(p^2)$ space complexity, where p is the dimensionality of the parameters to estimate.

To avoid computing the expensive Jacobian and Hessian matrices, SDM proposes to employ a generic descent map - $\mathbf{R} \in \mathbb{R}^{2L\times 128L}$ and $\mathbf{b} \in \mathbb{R}^{2L}$ to represent all the face images' $-2\mathbf{H}_f^{-1}\mathbf{J}_h^{\mathrm{T}}$ and $-2\mathbf{H}_f^{-1}\mathbf{J}_h^{\mathrm{T}}h(\mathbf{x}_*, I)$. $\mathbf{R}$ and $\mathbf{b}$ define a linear mapping between $\Delta\mathbf{x}$ and $h(\mathbf{x}_0, I)$, which can be solved as follows:

$$\underset{\mathbf{R},\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| \Delta\mathbf{x}_*^i - \mathbf{R}h\big(\mathbf{x}_0^i, I_i\big) - \mathbf{b} \right\|_2^2 \tag{2.5}$$

where N is the number of training images and $\Delta\mathbf{x}_*^i = \mathbf{x}_*^i - \mathbf{x}_0^i$. Since it is difficult to approach the target shape with a single update step, a sequence of descent maps denoted as $\{\mathbf{R}_k\}$ and $\{\mathbf{b}_k\}$ are learned during training. Then, for a new face image, in each iteration k, the shape increment can be calculated as:

$$\Delta\mathbf{x}_k = \mathbf{R}_k h(\mathbf{x}_{k-1}, I) + \mathbf{b}_k \tag{2.6}$$

Using the simple but effective supervised setting mentioned above, SDM converts the nonlinear optimization problem of face alignment into a linear least

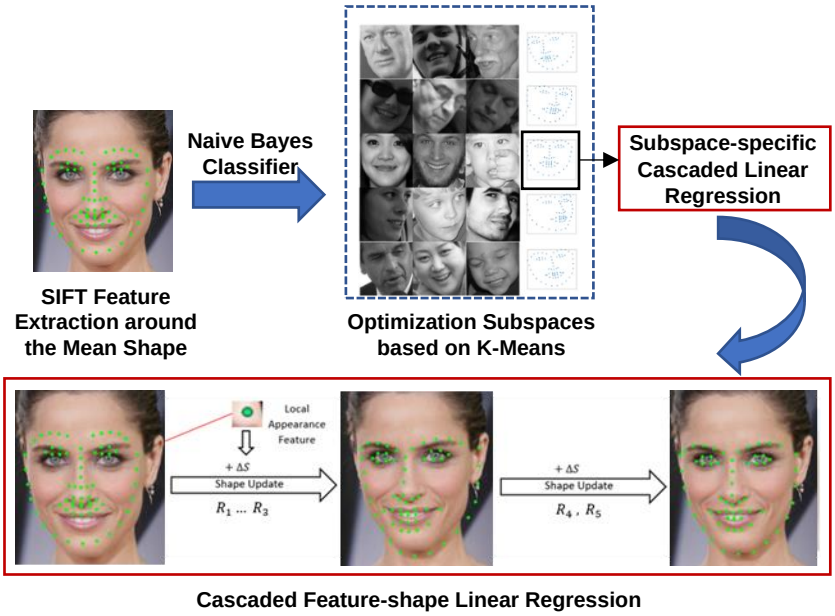


Figure 2.2: The workflow of MS-SDM.

squares problem, which can be solved efficiently with a closed-form solution. However, like the Newton's method, SDM is sensitive to the initial points and prone to local optima. It can also be found that the feature extraction function $h(\mathbf{x}, I)$ is parameterized not only by landmark positions $\mathbf{x}$ but also by the face image I. Different face images' $-2\mathbf{H}_f^{-1}\mathbf{J}_h^{\mathrm{T}}$ and $-2\mathbf{H}_f^{-1}\mathbf{J}_h^{\mathrm{T}}h(\mathbf{x}_*, I)$ can thus hardly be represented with a generic descent map $\mathbf{R}$ and $\mathbf{b}$. As a result, one optimization process as shown with Eq. 2.5 is insufficient to address the unconstrained face alignment problem which covers a wide range of facial shapes and appearances.

2.3.2 Multi-subspace SDM

To tackle the limitations of SDM, I propose a novel two-step method – MS-SDM (see Fig. 2.2). MS-SDM first applies k-means to partition the original optimization space of face alignment (see Eq. 2.2) into different subspaces which show explicit semantic meanings. Then, for each subspace, an exclusive linear regression from facial appearance features to the shape increment is learned via SDM. In the testing phase, an image sample will first be assigned into the correct subspace with a

a. Subspaces learned from $\Delta\mathbf{x}$ b. Subspaces learned from Δh

Figure 2.3: Comparison between the subspaces learned from $\Delta\mathbf{x}$ and Δh. Each row represents a training subset which contains three example images and the mean shape of all the faces in the subset. The k-means cluster amount was set as 5.

pre-trained Naive Bayes classifier. The corresponding feature-shape regression will later be called to predict the shape increment:

$$\Delta\mathbf{x}_k = \mathbf{R}_{k,s}h(\mathbf{x}_{k-1}, I) + \mathbf{b}_{k,s} \tag{2.7}$$

where s represents the subspace label.

1) Optimization Subspace Learning via K-means

As shown in Eq. 2.4, data samples with a similar regression target $\Delta\mathbf{x}$ will be more likely to fall inside the same optimization space and have compatible descent maps. I hence apply the classic clustering algorithm - k-means on all the training samples' $\Delta\mathbf{x}$ to find out the principal facial shape variations and divide the original training set into several subsets. In order to leverage the useful shape information, the original $\Delta\mathbf{x}$ of each training sample is used during the clustering process. As shown in Fig. 2.3a, the training subsets generated in this way show a high correlation with the head pose. Specifically, each subset relates to a particular head pose such as left-profile face, right-profile face, left-rolling face and right-rolling face.

During the optimization process of face alignment, the target facial shape increment $\Delta\mathbf{x}$ is predicted from the ground-truth feature deviation $\Delta h = h(\mathbf{x}_0, I) - h(\mathbf{x}_*, I)$ (see Eq. 2.4). This indicates that data samples having similar Δh will tend to share the same descent map $\mathbf{R}$ and $\mathbf{b}$. Inspired by this insight, I further apply k-means on the training samples' Δh to partition the optimization space. Similar to the subspaces learned from $\Delta\mathbf{x}$, the subspaces generated this time also show a clear semantic link to the head pose (see Fig. 2.3b).

2) Robust Subspace Prediction with a Naive Bayes Classifier

The subspace learning process proposed above depends on the ground-truth facial shape which is unknown during testing. The main challenge in shape prediction at the testing stage now turns into assigning an image sample into the corresponding subspace correctly. A straightforward solution of this problem is to apply a multiclass classifier (e.g. Random Forest, SVM or Naive Bayes) which learns the subspace label from the facial appearance feature. In this work, I adopt the Naive Bayes classifier as the subspace classifier since it inherently considers the relative proximity between the data sample and the subspace when making the prediction. Specifically, the Naive Bayes classifier predicts a class label $\hat{y} = C_k$ with the following optimization process:

$$\hat{y} = \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}}\, p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{2.8}$$

where $x_1, \cdots, x_n$ represent the features of a data sample and are assumed to be conditionally independent from each other. $p(C_k)$ is the prior probability of class C_k, and $p(x_i|C_k)$ is formulated with a Gaussian distribution to represent the likelihood of observing feature x_i with class label C_k. The overall likelihood $\prod_{i=1}^{n} p(x_i|C_k)$ can be viewed as a measure of the distance between the data sample and the class centre. If the data sample is far away from the class centre, then $\prod_{i=1}^{n} p(x_i|C_k)$ is small, otherwise, $\prod_{i=1}^{n} p(x_i|C_k)$ is large. Since $\prod_{i=1}^{n} p(x_i|C_k)$ directly contributes to the optimization process of predicting the subspace label,

the relative proximity between the sample and the subspace is then naturally embedded into the Naive Bayes Classifier. This can avoid assigning a sample into an incompatible subspace.

In the testing phase, I first place a mean-shape into the face bounding box. Then I extract the SIFT feature around each landmark of the mean-shape from the face image (see Fig. 2.2) and concatenate all the features together to form an appearance feature vector to feed into the Naive Bayes classifier.

2.4 Experiments

Dataset. I evaluate the proposed MS-SDM on a widely-applied benchmark dataset – 300W (Sagonas et al., 2013) and the NTHU Drowsy Driver Detection (NTHU-DDD) video dataset (C. H. Weng et al., 2016). 300W is a compilation of five challenging face image datasets, including AFW (X. Zhu & Ramanan, 2012), LFPW (Belhumeur et al., 2013), HELEN (Le et al., 2012), XM2VTS (Messer et al., 1999) and IBUG. The face images in 300W cover a wide range of head poses, facial expressions and appearance, and each image is provided with 68 hand-annotated landmarks. To train the MS-SDM model, I use 3,148 images which are composed of the whole AFW set and the training sets of LFPW and HELEN. The full testing set contains 689 images in total and can be divided into two parts – a common set (554 images) and a challenging set (135 images). The common set is composed of the testing sets of LFPW and HELEN. Since IBUG mainly consists of face images with large head poses and extreme facial expressions, it is treated as the challenging set. To eliminate the face detection's influence on the final results, I use the ground-truth face bounding boxes provided by 300W in the experiment.

Evaluation Metric. The widely-accepted point-to-point root mean square error (normalised by the inter-pupil distance) between the detected facial landmarks and the ground-truth annotations is used to measure the face alignment error. For simplicity, the '%' is omitted in the results presented below.

Implementation Details.

a) I adopt the data augmentation method proposed in (Xiong & De la Torre, 2013) to enlarge the training set and improve the landmark detection model's generalization capability. Specifically, for each training image, I perturb its original face bounding box 10 times, each time applying a random translation and scale for perturbation.

b) For learning the optimization subspaces, I empirically set the k-means cluster amount as 5, which produces promising results in my experiment.

c) Instead of using a generic mean shape for initialization, I utilize the mean shape calculated from all the training samples falling inside the specific subspace to initialize the corresponding shape regression.

d) When training the Naive Bayes classifier, I found that using the facial appearance features extracted with multiple initial shapes delivers a higher classification accuracy comparing to using the features extracted with a single initial shape. I hence feed the image features extracted with all the subspace-specific mean shapes into the Naive Bayes classifier.

e) For fair comparison, I re-implement SDM (Xiong & De la Torre, 2013) and GSDM (Xiong & De la Torre, 2015) by myself, and train the corresponding face alignment models with the same data set as that used for training MS-SDM. My implementations achieve a similar landmark detection accuracy as that reported in other mainstream implementations (Z. Zhang et al., 2014).

f) In my experiment, the training and testing sets are the same as those used in a benchmark test (Sagonas et al., 2013). The test has become a widely-accepted standard for evaluating face alignment methods in the research community. To fairly compare the proposed MS-SDM with the other methods, I apply that benchmark test, train MS-SDM models on the aforementioned training set, and compute the alignment error on the three testing sets. In this way, it's sufficient to evaluate the method's robustness

Table 2.1: The landmark detection error of SDM, GSDM and MS-SDM on the testing set.

	Common Set	**Challenging Set**	**Full Set**
SDM	5.59	15.38	7.51
GSDM	5.39	12.57	6.80
MS-SDM	**5.30**	**12.29**	**6.47**

Figure 2.4: Example visual results of MS-SDM on the testing set.

/generalization ability, hence no other cross-validation approaches are applied.

2.4.1 Comparison with SDM

As shown in Table 2.1, the proposed MS-SDM outperforms SDM (Xiong & De la Torre, 2013) on all the testing sets, especially on the challenging set. The challenging set consists of faces with various head poses and facial expressions, which implies that multiple descent maps are needed for accurate facial landmark detection. However, SDM only learned a single generic descent map which is prone to generating mistaken descent directions for some face images in the challenging set, especially for those with extremely large head pose and facial expression. On the contrary, MS-SDM has learned multiple feature-shape regressions (or descent maps). With a correct subspace prediction, the face image will have a much bigger chance to be assigned with an appropriate descent map. To further demonstrate the robustness of the proposed MS-SDM, I present some of its visual results on the testing set in Fig. 2.4.

2.4.2 Comparison with GSDM

As introduced in the Related Work section, GSDM (Xiong & De la Torre, 2015) develops a different optimization space (or domain) partition method for improving SDM, which has demonstrated its effectiveness in real-time 2D facial tracking in a video. For selecting the correct optimization subspace during testing, GSDM requires an estimation of the facial shape which should be very close to the ground-truth. This is infeasible in single image face alignment. To enable comparison between MS-SDM and GSDM, I assume that the ground-truth shape of each face image is known beforehand and utilize it to do the optimization subspace selection in GSDM. Following (Xiong & De la Torre, 2015), I partition the optimization space of the face alignment problem into eight subspaces for both MS-SDM and GSDM. Within each subspace, a linear feature-shape regression is learned via SDM using the face images falling inside the subspace. As shown in Table 2.1, MS-SDM delivers a higher landmark detection accuracy than GSDM on all the testing sets. It is also worth pointing out that MS-SDM can be applied to both the face alignment on a still image and the facial tracking in a video.

2.4.3 Real-time 2D Facial Tracking Results

I further test the proposed MS-SDM on real-time 2D facial tracking in a video. I train a tracking model of MS-SDM with the face images from 300W and Multi-PIE (R. Gross et al., 2010), and test the model on a benchmark video set and live video streams. Fig. 2.5 shows some visual tracking results of the proposed MS-SDM on the NTHU-DDD video dataset (C. H. Weng et al., 2016). The tracked facial landmarks can be used for analysing the driver's physical status such as the drowsiness for safe driving alert. I also ported the tracking model to the mobile device for implementing an Android real-time facial tracking application. As shown in Fig. 2.6, the application can robustly track the user's face spanning a wide range of head poses and facial expressions. It can be further incorporated into

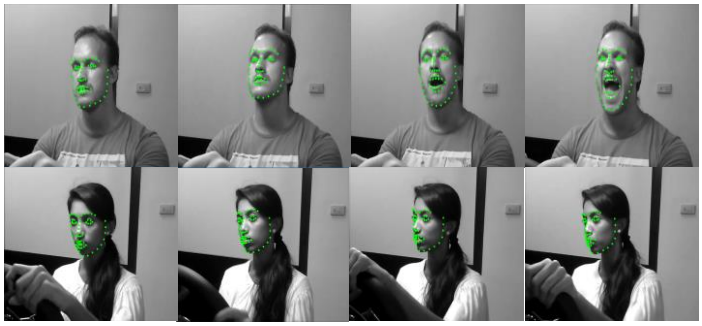


Figure 2.5: Tracking results on the NTHU Drowsy Driver Detection (NTHU-DDD) video dataset.

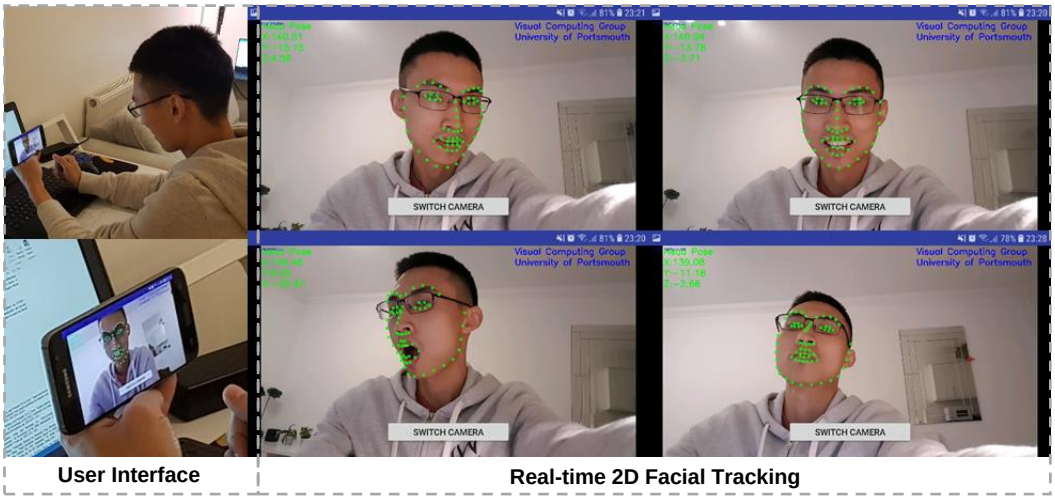


Figure 2.6: Screenshots of the MS-SDM-based facial tracking mobile application.

a variety of mobile consumer applications such as those for automated face makeup and personalised emoji generation. The experimental results intuitively demonstrate the potential of MS-SDM in real-world facial tracking applications.

2.5 Conclusion

Even with a very elegant formulation, SDM achieves the state-of-the-art face alignment performance. However, SDM is a local algorithm and prone to generating mistaken descent directions for face images with large head pose and facial expression. To tackle the limitations of SDM, this chapter proposed a novel two-step method – MS-SDM, which can push SDM closer to unconstrained face

alignment. By applying k-means on facial appearance features, MS-SDM finds the optimization subspaces of face alignment which exhibit a clear semantic link to the head pose. For each subspace, MS-SDM learns an exclusive feature-shape regression via SDM. Then, with a robust Naive Bayes classifier, each testing sample can be allocated with an appropriate subspace shape regression, which reduces the risk of generating incorrect descent directions that severely degrade the landmark detection accuracy. The proposed MS-SDM has been tested on challenging face datasets and a mobile facial tracking application. It showed improved face alignment performance comparing against SDM and its variant - GSDM. However, it can be found that MS-SDM has not exploited the complementary information between different subspaces, which might limit its performance. As a future work, I will explore to combine subspace-specific shape regressions via compositional learning (S. Zhu et al., 2016) to further improve the facial landmark detection accuracy.

Chapter 3
Real-time 3D Facial Tracking via Cascaded Compositional Learning

Foreword

This chapter extends the facial performance capture from sparse 2D facial shape to dense 3D facial geometry. Specifically, it develops a novel and robust facial motion regression method for real-time 3D facial tracking from a monocular RGB camera. The proposed method adopts a similar cascaded regression framework as that applied in the previous chapter, but extends the regression from unimodal 2D facial landmarks to multi-modal 3D facial motion parameters. It achieves the state-of-the-art 3D facial tracking performance with very low data and computational cost, hence providing a highly elegant and practical solution for capturing the 3D face in real-world applications. The chapter also deeply investigates the effect of synthesized facial images on training the regression model. It opens up a new perspective of incorporating the synthetic data to train non-deep learning methods, which can benefit a variety of facial image analysis tasks where the labelled real data is difficult to acquire.

The chapter is based on a journal paper that has been submitted to *IEEE Transactions on Image Processing* and is currently under review:

- **Lou, J.**, Cai, X., Dong, J., & Yu, H. (2020). Real-time 3D facial tracking via cascaded compositional learning. *IEEE Transactions on Image Processing*, under review.

In the paper, I proposed the facial motion regression method, implemented the corresponding 3D facial tracking interface with C++, did the experimental validation and wrote up the paper.

3.1 Introduction

Tracking 3D facial motion from a monocular RGB camera is a fundamental task which benefits a wide range of applications such as facial animation (C. Cao et al., 2014, 2015; C. Cao, Weng, Lin, et al., 2013), facial reenactment (Thies, Zollhofer, et al., 2016) and emotion recognition (Chen et al., 2015). Over the past years, a number of novel tracking algorithms (Y. Guo, Zhang, Cai, Cai, et al., 2018; Y. Guo, Zhang, Cai, Jiang, et al., 2018; Laine et al., 2017; Ma & Deng, 2019; McDonagh et al., 2016; Saito et al., 2016; C. Wang et al., 2016; Yoon et al., 2019) have been proposed, which led to rapid progress in this area. In particular, machine learning-based approaches that directly learn a regression function from image features to motion parameters greatly improve the tracking performance in speed, robustness and ease of use by circumventing the compute-intensive online optimization steps and leveraging a high-quality training corpus. Whereas the current state-of-the-art can deliver impressive tracking results even for very challenging cases such as large facial pose (C. Cao et al., 2014; Y. Guo, Zhang, Cai, Jiang, et al., 2018) and severe occlusion (Saito et al., 2016), the regression algorithms they applied are still not effective in dealing with multi-modal motion parameters.

Facial motion parameters such as those of head pose and expression vary significantly in scale and have different influences on facial geometry. Accurately regressing to multi-modal motion parameters from image features is a challenging task. Its learning process is prone to focusing more on parameters (e.g. 2D landmark displacements) with higher dimensionality and larger magnitude, while neglecting those (e.g. rotation angles) that impact heavily on facial geometry but with smaller magnitude. To solve this problem, previous methods either carefully chose weights to balance the parameter effects on feature selection when training a boosted ferns (McDonagh et al., 2016; Y. Weng et al., 2014) or minimized a more complicated photo-geometric difference loss instead of the parameter difference loss when training a convolutional neural network (Y. Guo, Zhang, Cai,

Cai, et al., 2018; Y. Guo, Zhang, Cai, Jiang, et al., 2018). For the first method, the process of finding appropriate weights is somewhat clumsy and it's arguable if those empirical weights can correctly reflect the parameter's significance. The latter method embeds parameter effects into the gradient of the loss function but at high computational cost.

To tackle the aforementioned problems, this chapter adopts the compositional learning framework and proposes a novel boosting method that can efficiently cope with the modality variety in output variables. The proposed method first learns a modular boosted ferns (X. Cao et al., 2014) which is a shallow composition of several independent regression models with each is a boosted ferns trained targeting only partial output variables of the same modality. All fern leaves are then simultaneously optimized by minimizing a global loss function defined on all output variables, which can be solved efficiently with Ridge Regression (Hoerl & Kennard, 1970). The complementary information between the old biased ferns is thus injected into the refined fern leaves, producing a new boosted ferns which is a deep composition of the pre-learnt modality-specific regression models and has much stronger predictive power. The method is named as Globally-optimized Modular Boosted Ferns - GoMBF. As in (X. Cao et al., 2014; Dollár et al., 2010), the chapter then builds facial motion regression with a cascade of GoMBFs (GoMBF-Cascade) which progressively update motion parameters from an initial state by calling GoMBF to estimate an increment stage-by-stage. Extensive experiments on in-the-wild videos demonstrate that GoMBF is superior in both the fitting power and the learning speed comparing against the traditional boosted ferns that has been widely applied in 2D/3D facial shape regression(C. Cao et al., 2014; C. Cao, Weng, Lin, et al., 2013; X. Cao et al., 2014). The resulting GoMBF-Cascade regression delivers competitive 3D facial tracking performance comparing to the state-of-the-art methods (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Ma & Deng, 2019) which require much more training data or have a much higher computational complexity.

Along with a reliable regression algorithm, quality training data is another key factor to the tracking model's robustness. For 3D facial tracking, the training data typically means facial images paired with the ground truth 3D geometry. Such data is normally acquired by multi-view stereo (Laine et al., 2017), photometric (Y. Guo, Zhang, Cai, Jiang, et al., 2018) or 2D landmark-based (C. Cao et al., 2014) reconstruction which requires either complicated and expensive multi-camera setups or laborious manual annotations. Alternatively, synthetic generation of training imagery provides a more economic and efficient data collection way. This approach has shown effectiveness on training deep convolutional neural networks for accurate 3D facial tracking and reconstruction in recent studies (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2016, 2017). However, it remains unclear whether the synthetic data also works on training non-deep learning methods such as GoMBF-Cascade. This chapter explores this question via progressively adjusting the naturalness of synthetic images for training GoMBF-Cascade and comparing between tracking models that are trained on real data, on synthetic data and on a mixture of data. The experimental results show that the GoMBF-Cascade models trained purely on synthesized images have poor tracking performance on real videos and become more biased after incorporating the synthetic data into training.

In summary, the main contributions of this chapter are as follows:

i) Based on compositional learning, a novel boosting algorithm – GoMBF is developed. It deals effectively with the modality variety in output variables. GoMBF shows stronger fitting power and a faster learning speed when comparing with the conventional boosted ferns (C. Cao et al., 2014; C. Cao, Weng, Lin, et al., 2013; X. Cao et al., 2014). It can be seamlessly adapted to any other multi-output regression tasks in theory.

ii) By cascading GoMBFs for facial motion regression, it achieves a competitive 3D facial tracking performance compared with the state-of-the-art methods (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Ma & Deng, 2019), which rely

on large-scale training data or bear much higher computational complexity. It thus offers a robust and very practical solution to real-time 3D facial tracking.

iii) It carries out an in-depth investigation into the effect of synthetic data on training GoMBF-Cascade for 3D facial tracking, which provides a novel view of the synthetic data's role in training non-deep learning method for facial image analysis where the real labelled data is difficult to obtain.

3.2 Related Work

Real-time 3D facial motion capture from a monocular RGB video has been extensively studied in computer graphics and vision communities. It is normally achieved by estimating a group of parameters which encode facial expression and head pose within a low-dimensional space from video frames. Generally, there are two types of approaches to estimate those parameters - optimization-based approach and learning-based approach, which divides the existing studies into two main streams. This section reviews the most relevant works from the two categories and also discusses how the synthetic data has been used in learning-based approaches. For a more comprehensive review on related topics, interested readers are directed to (Zollhöfer et al., 2018).

3.2.1 Optimization-based Approaches

Optimization-based approach is built upon the idea of analysis-by-synthesis where a parametric face model is iteratively adapted until the synthesized face matches the target image. It is formulated as minimizing a highly non-linear objective function which typically enforces alignment on sparse/dense feature points (C. Cao, Weng, Lin, et al., 2013; Jeni et al., 2015; C. Wang et al., 2016) and pixel intensities (Ma & Deng, 2019; Thies, Zollhofer, et al., 2016) between the synthesized result and the input data, while regularizing the estimated shape parameters to lie within a valid range for generating a plausible face. Solving this optimization problem usually requires massive computing power such as GPU

acceleration to achieve real-time performance (Ma & Deng, 2019; Thies, Zollhofer, et al., 2016). This hinders the approach's deployment to platforms with limited computing resources.

3.2.2 Learning-based Approaches

Learning-based approach bypasses the costly optimization step by estimating facial motion parameters from image features through a regression learned from a hand-picked training corpus. Cao et al. (C. Cao et al., 2014; Y. Weng et al., 2014) pioneered this area by employing a two-level boosted regression – Explicit Shape Regression (ESR) (X. Cao et al., 2014) to map facial appearance features to motion parameters. Their method was trained on public image datasets with estimated 3D facial data and achieved impressive tracking performance on in-the-wild videos. The work opened up a new era of learning-based 3D facial tracking and motivated a bunch of follow-ups (C. Cao et al., 2015; McDonagh et al., 2016; Saito et al., 2016) which extended the tracking to more challenging cases such as capturing facial geometry details (e.g. wrinkles and dimples) (C. Cao et al., 2015) and tracking under severe occlusions (Saito et al., 2016). Despite the great success achieved by these works, the boosted ferns employed in ESR is deficient in handling the modality variety of motion parameters whose scale and influence on facial geometry differ a lot from each other. To mitigate this problem, a few studies (McDonagh et al., 2016; Y. Weng et al., 2014) applied a weighting-vector to balance the parameter effects on feature selection in fern learning. This intuitive strategy is moderately inefficient and it's doubtful if those empirical weights can fully reflect the parameter's significance. More recent studies (Y. Guo, Zhang, Cai, Cai, et al., 2018; Y. Guo, Zhang, Cai, Jiang, et al., 2018) instead employed a deep convolutional neural network coupled with a photo-geometric difference loss to learn the facial motion regression. This method inherently incorporates the motion parameter's influence on facial geometry into the gradient of the loss function, which however bears a high computational complexity. Alternatively, the proposed GoMBF first learns an exclusive boosted ferns for each kind of motion

parameters and then optimizes all fern leaves towards the whole regression target with linear regression, which explicitly handles the output variable's modality variety in a fairly efficient manner.

3.2.3 Learning from Synthetic Data

In contrast to traditional 3D facial data harvesting methods which need multi-camera setups or manual annotations, synthetic generation of training imagery offers a highly efficient and economic data collection way. Learning from synthetic data is attracting more and more attention in 3D facial tracking and reconstruction (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2016, 2017). Richardson et al. (Richardson et al., 2016) proposed to render photo-realistic 3D facial meshes and images using 3D Morphable Model (3DMM) (Paysan et al., 2009) and Phong illumination (Phong, 1975) for training a convolutional neural network (CNN) for 3D face reconstruction. Though the network was trained purely on synthetic data, it generalized well to real-world face images. Guo et al. (Y. Guo, Zhang, Cai, Jiang, et al., 2018) later used albedo and lighting coefficients inferred from real face images to render more natural-look faces for training the CNN. Their model achieved high-quality tracking results on in-the-wild videos. A more recent study (Kortylewski et al., 2018) shows that priming deep networks by pre-training them with synthetic faces is helpful, e.g. it can reduce the negative effects of the training data bias. Whereas there is continuous evidence manifesting that the synthesized faces favour deep learning methods, it remains unclear if such data also benefits non-deep learning methods. To my knowledge, only McDonagh and his colleagues (McDonagh et al., 2016) have succeeded in learning a boosted ferns from the synthesized faces for personalized 3D facial tracking. However, their synthetic generation of training imagery was based on a high-quality facial rig of the user's face obtained from an offline capture system and a simulated illumination driven by light probe data acquired at the target environment. This process can hardly be adapted to unconstrained facial tracking where the target environment is unknown in the

training phase. This chapter provides a novel view of the synthetic data's role in training non-deep learning methods by incorporating three kinds of synthetic data for training GoMBF-Cascade and comparing tracking models trained on real data, on synthetic data and on a mixture of data.

3.3 Method Overview

This section overviews the developed 3D facial tracking framework. It first introduces the parametric face model for representing the facial shape, then formulates the tracking workflow which is driven by the proposed GoMBF-Cascade motion regression.

3.3.1 Parametric Face Model

A 3D facial mesh is typically formed with a vector of stacked vertex coordinates $S = [x_1, y_1, z_1, \dots, x_n, y_n, z_n]^T$ ($n = 53{,}215$ in this study) and a predefined connectivity. The lengthy coordinate vector can be calculated as a weighted sum of a few basis vectors, which leaves weights the only control parameters and generates a low-rank representation of the facial mesh:

$$S = B_{id}\boldsymbol{\alpha} + B_{exp}\boldsymbol{\delta} \tag{3.1}$$

As shown in Eq. 3.1, $B_{id} = \left[\mathbf{b}_0^{id}, \mathbf{b}_1^{id} \dots, \mathbf{b}_{m_{id}}^{id}\right]$ is the linear basis for representing facial identity, in which $\mathbf{b}_0^{id}$ is the mean face in neutral expression. $\boldsymbol{\alpha} = \left[1, \alpha_1, \dots, \alpha_{m_{id}}\right]^T$ denotes the relevant identity coefficients. $B_{exp} = \left[\mathbf{b}_1^{exp}, \dots, \mathbf{b}_{m_{exp}}^{exp}\right]$ is composed of delta blendshapes of the mean face $\mathbf{b}_0^{id}$ for representing facial expression, whose coefficients - $\boldsymbol{\delta} = \left[\delta_1, \dots, \delta_{m_{exp}}\right]^T$ are bounded between 0 and 1. I get B_{id} ($m_{id} = 80$ as only the first 80 principal components are used in this study) from the Basel Face Model (BFM) (Paysan et al., 2009) and generate B_{exp} ($m_{exp} = 46$) from FaceWarehouse (C. Cao, Weng, Zhou, et al., 2013) using deformation transfer (Sumner & Popović, 2004).

To map S which is measured in world space to image space, I apply an ideal pinhole camera model. Given a 3D point $\mathbf{v}$ in S, the corresponding 2D image point $\mathbf{p} = [p_x, p_y]^T$ can be obtained as:

$$\mathbf{p} = \mathbf{\Pi_Q}(\mathbf{Rv} + \mathbf{t}) \tag{3.2}$$

where $\mathbf{R}$ is the rotation matrix parameterized by Euler angles (*yaw*, *pitch* and *roll*) $\boldsymbol{\theta} \in \mathbb{R}^3$, and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector. $\mathbf{\Pi_Q}$ denotes a perspective projection operator parameterized by $\mathbf{Q} = [f, 0, u_0; 0, f, v_0; 0,0,1]$ in which f is the focal length and (u_0, v_0) is the image centre. In practice, the estimated 3D face - S and the camera model - $\{\mathbf{Q}, \mathbf{R}, \mathbf{t}\}$ may not fully match the face image. To compensate for this discrepancy, I follow (C. Cao et al., 2014) by using a 2D landmark displacement vector $\mathbf{D} \in \mathbb{R}^{132}$ to add onto the projected landmark coordinates to acquire 66 more accurate landmarks on the image.

The combination of parameters - $\{\boldsymbol{\alpha}, \boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\theta}, \mathbf{t}, \mathbf{D}\}$ provides a compact representation of both the 3D and 2D facial shapes. $\boldsymbol{\alpha}$ and $\mathbf{Q}$ are invariant across the whole video sequence for the same human subject. $P = [\boldsymbol{\delta}; \boldsymbol{\theta}; \mathbf{t}; \mathbf{D}] \in \mathbb{R}^{184}$ controls facial motion and changes frame by frame.

3.3.2 Tracking Workflow

Based on the parametric face model, 3D facial tracking from a monocular RGB video can be casted into regressing motion parameters P from a video frame I (see Fig. 3.1):

$$P = \mathcal{R}(I, \boldsymbol{\alpha}, \mathbf{Q}, P^0) \tag{3.3}$$

where $\mathcal{R}(\cdot)$ is the regression function, P^0 denotes the initial motion parameters generated from the previous frame's estimation for enforcing temporal coherence. I build $\mathcal{R}(\cdot)$ by learning a linear sequence of GoMBFs (GoMBF-Cascade) which gradually refines P from P^0 to fit with the current frame. $\boldsymbol{\alpha}$ and $\mathbf{Q}$ are estimated from the first frame and keep fixed for the remaining frames.

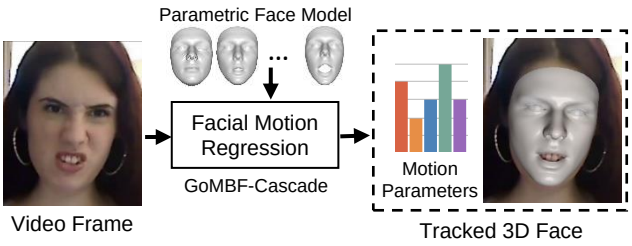


Figure 3.1: 3D facial tracking workflow.

3.4 Facial Motion Regression with GoMBF-Cascade

This section starts with introducing boosted ferns (X. Cao et al., 2014) which is a key building block of the whole facial motion regression method. Then it elaborates the proposed globally-optimized modular boosted ferns - GoMBF and GoMBF-Cascade regression.

3.4.1 Boosted Ferns

Prediction. Boosted ferns (X. Cao et al., 2014) is an ensemble of ferns, each fern addresses the residual of the regression target left by the preceding ferns. Its prediction is therefore the sum of all ferns' outputs. Fern is a particular instance of decision tree, which applies an identical node-splitting test for all nodes at the same tree level. The prediction of a fern with $F + 1$ levels can be formulated in a compact form:

$$\mathbf{y} = \mathbf{w}\phi(\mathbf{x}) \tag{3.4}$$

where $\mathbf{w}$ is a matrix of 2^F columns with each column stores a leaf node's prediction of output variables, $\phi(\cdot)$ represents the fern's structure (the learned node-splitting tests) which maps the data sample $\mathbf{x}$ to a one-hot vector of 2^F rows with each row indicating if $\mathbf{x}$ falls inside a leaf node or not (1 for yes, 0 for no), and $\mathbf{y}$ is the fern's prediction of $\mathbf{x}$. The prediction of a boosted ferns (see Fig. 3.2a) with K ferns is thus:

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{w}_i \phi_i(\mathbf{x}) = W\Phi(\mathbf{x}) \tag{3.5}$$

where $W = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ and $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}); \ldots; \phi_K(\mathbf{x})]$ which is highly sparse.

Training. Training a boosted ferns equals to progressively training a sequence of ferns, where each fern's training loss is defined on the residual of the regression target. Specifically, a fern with $F + 1$ levels is built in two consecutive steps:

a) Learn the mapping function - $\phi(\cdot)$. It is to learn a series of node-splitting tests, each for sending a data sample $\mathbf{x}$ to the right child node if the test is satisfied or to the left child node if not. Typically, a node-splitting test is about selecting a feature from $\mathbf{x}$ and comparing it to a threshold. As in (X. Cao et al., 2014), I calculate the differences (referring to image pixel differences in my case) between $\mathbf{x}$'s elements and select the one that has the highest Pearson Correlation with a random projection (generated from a Gaussian distribution) of the regression target as the feature for splitting the node. A threshold is then randomly sampled from a uniform distribution which is scaled by the selected feature's maximum absolute value in the training set (X. Cao et al., 2014). After repeating the process of feature selection and threshold sampling F times, the fern's $\phi(\cdot)$ can be obtained.

b) Learn the leaf matrix - $\mathbf{w}$. With the learned $\phi(\cdot)$, all training samples can be sent level by level from the fern root all the way down to one of the 2^F leaf nodes. For each leaf node, I acquire its prediction of output variables by averaging the regression targets of all training samples falling inside this node with a shrinkage to overcome overfitting (X. Cao et al., 2014) and save it into the corresponding column of $\mathbf{w}$.

3.4.2 Globally-optimized Modular Boosted Ferns

Whereas boosted ferns has been successfully applied in 2D/3D shape regression (C. Cao et al., 2014; C. Cao, Weng, Lin, et al., 2013; X. Cao et al., 2014; McDonagh et al., 2016; Y. Weng et al., 2014), I found it has limitations when regressing to multi-modal output variables such as facial motion parameters $P = [\boldsymbol{\delta}; \boldsymbol{\theta}; \mathbf{t}; \mathbf{D}]$.

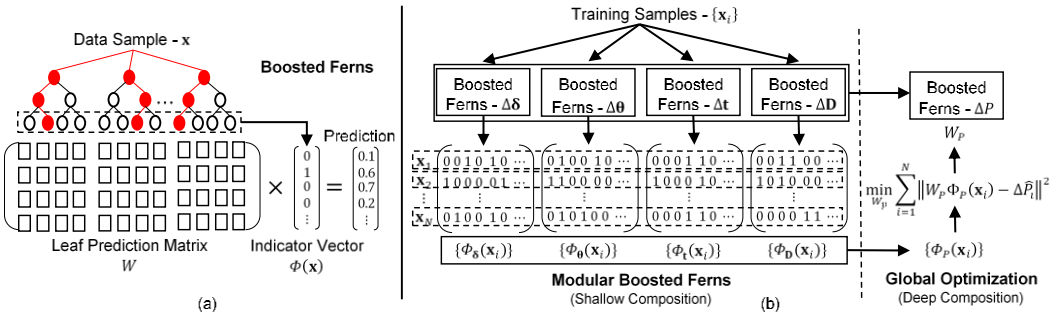


Figure 3.2: Illustration of the boosted ferns and the GoMBF built with compositional learning: (a) boosted ferns, (b) GoMBF.

As shown above, the prediction of a boosted ferns relies heavily on the node-splitting features. The aforementioned correlation-based feature selection method can efficiently learn good features when output variables are of a single modality. However, if output variables such as motion parameters contain multiple modalities, it is prone to selecting features that are more discriminative to output variables (e.g. 2D landmark displacements $\mathbf{D}$) with higher dimensionality and larger magnitude, while less informative to those (e.g. rotation angles $\boldsymbol{\theta}$) which are relatively negligible in numerical scale but significant in semantics. This severely degrades boosted fern's fitting power. To solve the problem, I follow the compositional learning framework and propose a globally-optimized modular boosted ferns – GoMBF, which is built in two consecutive phases (see Fig. 3.2b): 1) learn a modular boosted ferns in which each module regresses partial output variables of the same modality; 2) optimize all fern leaves towards the whole regression target by solving a linear regression.

1) Learning a Modular Boosted Ferns

A modular boosted ferns is a shallow composition of multiple regression models with each is also a boosted ferns trained independently for regressing partial output variables of the same modality, which refers to the increment of a kind of motion parameters in my case (the number of ferns is $K_{\boldsymbol{\delta}}$, $K_{\boldsymbol{\theta}}$, $K_{\mathbf{t}}$, $K_{\mathbf{D}}$ respectively):

$$\begin{cases} [\Delta\boldsymbol{\delta}; \mathbf{0}; \mathbf{0}; \mathbf{0}] = W_{\boldsymbol{\delta}}\Phi_{\boldsymbol{\delta}}(\mathbf{x}) \\ [\mathbf{0}; \Delta\boldsymbol{\theta}; \mathbf{0}; \mathbf{0}] = W_{\boldsymbol{\theta}}\Phi_{\boldsymbol{\theta}}(\mathbf{x}) \\ [\mathbf{0}; \mathbf{0}; \Delta\mathbf{t}; \mathbf{0}] = W_{\mathbf{t}}\Phi_{\mathbf{t}}(\mathbf{x}) \\ [\mathbf{0}; \mathbf{0}; \mathbf{0}; \Delta\mathbf{D}] = W_{\mathbf{D}}\Phi_{\mathbf{D}}(\mathbf{x}) \end{cases} \tag{3.6}$$

where $\Delta\boldsymbol{\delta}$, $\Delta\boldsymbol{\theta}$, $\Delta\mathbf{t}$ and $\Delta\mathbf{D}$ denote the predictions of motion parameter increments, $\mathbf{0}$ represents the zero vector with variant rows for extending the left output vector to match P's size. $W_{\boldsymbol{\delta}}$, $W_{\boldsymbol{\theta}}$, $W_{\mathbf{t}}$, $W_{\mathbf{D}}$ represent the leaf matrices of motion parameters - $\boldsymbol{\delta}$, $\boldsymbol{\theta}$, $\mathbf{t}$ and $\mathbf{D}$ respectively, $\Phi_{\boldsymbol{\delta}}$, $\Phi_{\boldsymbol{\theta}}$, $\Phi_{\mathbf{t}}$, $\Phi_{\mathbf{D}}$ are the corresponding mapping functions, and $\mathbf{x}$ is the data sample. Equation 3.6 can be written in a more compact form:

$$\Delta P = W_P\Phi_P(\mathbf{x}) \tag{3.7}$$

where $\Delta P = [\Delta\boldsymbol{\delta}; \Delta\boldsymbol{\theta}; \Delta\mathbf{t}; \Delta\mathbf{D}]$, $W_P = [W_{\boldsymbol{\delta}}, W_{\boldsymbol{\theta}}, W_{\mathbf{t}}, W_{\mathbf{D}}]$, $\Phi_P(\mathbf{x}) = [\Phi_{\boldsymbol{\delta}}(\mathbf{x}); \Phi_{\boldsymbol{\theta}}(\mathbf{x}); \Phi_{\mathbf{t}}(\mathbf{x}); \Phi_{\mathbf{D}}(\mathbf{x})]$, and the subscript P denotes the composition of all motion parameters (for a detailed explanation of motion parameters, please refer to Section 3.3.1). The method reduces the original difficult regression task to four simpler sub-tasks which require a very small number of ferns for each sub-task and can be solved efficiently using parallel programming. This in turn avoids the interference from the output variable's modality variety on feature selection during fern learning.

2) Global Optimization

Due to the nature of modular boosted ferns, each module learns features biased towards partial output variables of a specific modality. Those features are complementary to each other, e.g. the features that are discriminative in estimating facial expression $\boldsymbol{\delta}$ could also benefit the estimation of 2D landmark displacements $\mathbf{D}$ as both parameters encode non-rigid facial motion. However, such complementary information between inter-modular ferns has not been exploited when making prediction in Eq. 3.7. For example, W_P's $W_{\boldsymbol{\delta}}$ only contributes to predicting expression coefficients $\boldsymbol{\delta}$, while remaining idle when

predicting the other three motion parameters. It makes the compositional regression model loosely articulated and less optimal. To this end, I propose to optimize all the pre-learnt fern leaves - W_P by minimizing a common objective function defined on the whole regression target:

$$\min_{W_p} \sum_{i=1}^{N} \left\| W_P \Phi_P(\mathbf{x}_i) - \Delta\widehat{P}_i \right\|^2 \tag{3.8}$$

where N is the number of training samples, $\Delta\widehat{P}_i$ is the regression target of sample i. Equation 3.8 is the well-known linear least squares problem which can be solved efficiently with Ridge Regression (Hoerl & Kennard, 1970). After updating Eq. 3.7 with the new W_p, I obtain a globally-optimized modular boosted ferns – GoMBF, which is a deep composition of the pre-learnt modality-specific regression models. In my 3D facial tracking experiments, GoMBF has shown stronger fitting power and a faster learning speed than the conventional boosted ferns (X. Cao et al., 2014). Moreover, it can be seamlessly applied to any other multi-output regression tasks in theory.

It is worth pointing out that GoMBF has conceptual links with two existing methods to some degree, which were developed for face alignment (Ren et al., 2014) and Random Forest refinement (Ren et al., 2015). However, GoMBF is fundamentally different from those two methods in the following two main aspects: i) GoMBF is designed to deal with the modality variety in regression output variables, while (Ren et al., 2014) is to learn discriminative local texture features for robust 2D landmark detection and (Ren et al., 2015) is to fill the gap between the training and the testing of Random Forest. ii) GoMBF is based on boosted ferns (boosting), while both (Ren et al., 2014) and (Ren et al., 2015) were based on Random Forest (bagging).

3.4.3 GoMBF-Cascade Regression

Following the basic idea of cascaded regression (X. Cao et al., 2014; Dollár et al., 2010; Ren et al., 2014) which has shown robustness in various shape regression

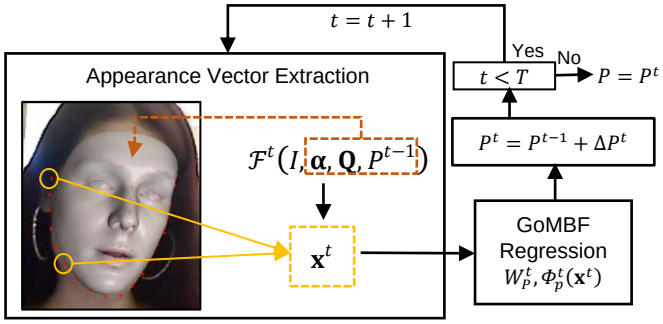


Figure 3.3: The pipeline of GoMBF-Cascade facial motion regression.

tasks, I frame the facial motion regression with a cascade of GoMBFs. The method is named as GoMBF-Cascade (see Fig. 3.3). For a video frame I, beginning with an initial motion vector P^0, GoMBF-Cascade gradually refines P by calling GoMBF to estimate a motion increment ΔP^t stage-by-stage:

$$P = P^0 + \sum_{t=1}^{T} \Delta P^t \tag{3.9}$$

$$\Delta P^t = W_P^t \Phi_P^t(\mathbf{x}^t)$$

$$\mathbf{x}^t = \mathcal{F}^t(I, \mathbf{\alpha}, \mathbf{Q}, P^{t-1})$$

where T is the number of stages, W_P^t and $\Phi_P^t(\cdot)$ represent the GoMBF learned at stage t. $\mathbf{x}^t$ is a vector of pixel intensities extracted from image I by $\mathcal{F}^t(\cdot)$ for representing the appearance of the facial shape - $\{\mathbf{\alpha}, \mathbf{Q}, P^{t-1}\}$ output from stage $t-1$. It can be found that Eq. 3.9 is a detailed expansion of Eq. 3.3. In the following, I will explain in detail the training, runtime prediction and appearance vector extraction of the GoMBF-cascade regression.

1) Training

To train the regression, I first create guess-truth motion parameter pairs $\{\hat{P}_i, P_{ij}^0\}$ for each training image I_i. The guess-truth pairs simulate the runtime situation where facial motion between two adjacent video frames is assumed to be small. Specifically, given a facial image I_i and its ground-truth facial motion parameters $\hat{P}_i$, I set the initial 2D landmark displacements as zeros and perturb

along $\hat{P}_i$'s three other dimensions - $\widehat{\boldsymbol{\delta}}_i, \widehat{\boldsymbol{\theta}}_i, \hat{\mathbf{t}}_i$ with random noise to get several guesses $\{P_{ij}^0\}$ of the initial motion parameters P_i^0:

- *Random Expression.* $P_{ij}^0 = [\boldsymbol{\delta}_{ij}^0; \widehat{\boldsymbol{\theta}}_i; \hat{\mathbf{t}}_i; \mathbf{0}]$, where $\boldsymbol{\delta}_{ij}^0 = \widehat{\boldsymbol{\delta}}_{i'}$ $(i' \neq i)$ is the ground-truth expression coefficients of image $I_{i'}$ which is randomly chosen from the training set.
- *Random Rotation.* $P_{ij}^0 = [\widehat{\boldsymbol{\delta}}_i; \boldsymbol{\theta}_{ij}^0; \hat{\mathbf{t}}_i; \mathbf{0}]$, where $\boldsymbol{\theta}_{ij}^0 = \widehat{\boldsymbol{\theta}}_i + \Delta\boldsymbol{\theta}_{ij}$. $\Delta\boldsymbol{\theta}_{ij}$ is composed of random Euler angles sampled from three independent normal distributions.
- *Random Translation.* $P_{ij}^0 = [\widehat{\boldsymbol{\delta}}_i; \widehat{\boldsymbol{\theta}}_i; \mathbf{t}_{ij}^0; \mathbf{0}]$, where $\mathbf{t}_{ij}^0 = \hat{\mathbf{t}}_i + \Delta\mathbf{t}_{ij}$. $\Delta\mathbf{t}_{ij}$ is a random translation vector whose elements are sampled from three independent normal distributions.

To find the appropriate number of guess-truth pairs, I have tested different values and visually compared the tracking performance of the trained models. For the random expression category, 15, 20, 25, 30, 35, 40 are tested; for the random rotation and translation categories, 4, 6, 8, 10, 12, 14 are tested. The experiment shows that the tracking performance improves as the number of guess-truth pairs increases at the beginning, then seems to have little improvement after the pair amount reaches a certain value. After testing, for each training image, I found generating 30 guess-truth pairs for the random expression category and 8 pairs for each of the other two categories is adequate for training a reliable tracking model.

After constructing the set of $\{I_i, \boldsymbol{\alpha}_i, \mathbf{Q}_i, \hat{P}_i, P_{ij}^0\}$, the GoMBF-cascade regression is trained in T stages. In each stage, I extract facial shape appearance vectors from all training images $\{I_i\}$ with a pre-built $\mathcal{F}^t(\cdot)$ and learn a GoMBF - $\{W_P^t, \Phi_P^t(\cdot)\}$ following the procedure explained in Section 3.4.1 and Section 3.4.2.

2) Runtime Prediction

For the first video frame, I locate the face using the Viola-Jones detector (Viola & Jones, 2004) and detect 66 landmarks with a pre-trained SDM (Xiong & De la Torre, 2013) model. Then, I predict its camera and facial shape parameters -

$\{\boldsymbol{\alpha}, \mathbf{Q}, P\}$ by fitting the aforementioned parametric face model to the detected 2D landmarks, which is achieved by minimizing the following energy with the coordinate-descent method:

$$E = E_{lan} + E_{reg} \tag{3.10}$$

$$E_{lan} = \sum_{k=1}^{66} \left\| \boldsymbol{\Pi}_{\mathbf{Q}} \left(\mathbf{R} \left(B_{id}\boldsymbol{\alpha} + B_{exp}\boldsymbol{\delta} \right)^{(l_k)} + \mathbf{t} \right) - \mathbf{p}_d^{(k)} \right\|^2$$

$$E_{reg} = w_1 \sum_{i=1}^{80} \left(\frac{\alpha_i}{\sigma_i} \right)^2 + w_2 \sum_{i=1}^{46} |\delta_i|$$

where E_{lan} represent the landmark fitting error and E_{reg} is the regularization term to enforce $\boldsymbol{\alpha}$ to stay statistically close to the mean and $\boldsymbol{\delta}$ to be sparse. In E_{lan}, $\mathbf{p}_d^{(k)}$ is the position of the kth detected 2D landmark and $\left(B_{id}\boldsymbol{\alpha} + B_{exp}\boldsymbol{\delta}\right)^{(l_k)}$ extracts the corresponding l_kth vertex on the 3D facial mesh. In E_{reg}, σ_i is the standard deviation of α_i, w_1 and w_2 balance the two sub-objectives. I set w_1 and w_2 as 10 and 1 respectively. For $\mathbf{Q}$, I set the focal length f as 1,000 and the principal point as the image center. This simple strategy is proven to be effective in my experiment. I then solve for $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ by applying the L-BFGS-B solver (Byrd et al., 1995) to constrain $\boldsymbol{\delta}$'s elements to lie within [0,1], and find the rigid facial motion $\{\mathbf{R}, \mathbf{t}\}$ using the POSIT algorithm (Dementhon & Davis, 1995). The energy converges in three iterations. After each iteration, I update the indices $\{l_k\}$ of contour landmarks on the facial mesh as in (X. Zhu et al., 2015). Once I had the estimations of $\{\boldsymbol{\alpha}, \boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\theta}, \mathbf{t}\}$, $\mathbf{D}$ can be obtained by subtracting the projected 2D landmark positions as computed in Eq. 3.2 from the detected 2D landmark positions.

For each subsequent frame, I initialize its motion parameter P based on the estimation $P_{prev} = \left[\boldsymbol{\delta}_{prev}; \boldsymbol{\theta}_{prev}; \mathbf{t}_{prev}; \mathbf{D}_{prev}\right]$ of the previous frame and call the learned GoMBF-Cascade regression to update P to align with the current facial shape. Specifically, I initialize P with $\boldsymbol{\theta}_{prev}$ and $\mathbf{t}_{prev}$, and set it's $\mathbf{D}$ as zeros. For facial expression, I found that directly inheriting $\boldsymbol{\delta}_{prev}$ for initialization will lead to implausible expression estimation and the error will accumulate across frames. This is probably due to the non-rigid nature of facial expression which makes the

distribution of expression coefficients complex and difficult to be covered by the training set with limited samples. To solve the problem, I select from the training set the expression coefficients that are closest to $\boldsymbol{\delta}_{prev}$ to initialize P. The distance between two expression coefficient vectors is measured as the mean average distance of landmarks extracted from the corresponding 3D facial meshes. In practice, I apply multiple initial Ps for regression and take the mean of all the outputs as the final prediction. Those initial Ps are generated with $\boldsymbol{\delta}_{prev}$'s L closest expression vectors in the training set. With the newly predicted facial motion parameters, I update the indices of contour landmarks on the facial mesh as in (X. Zhu et al., 2015).

3) Appearance Vector Extraction

As shown in Eq. 3.9, instead of directly sending the image into the regressor, a pixel intensity vector for representing facial shape appearance is extracted from the image by $\mathcal{F}^t(\cdot)$ and fed to the stage GoMBF - $\{W_P^t, \Phi_P^t(\cdot)\}$. The extracted pixels should contain the discriminative information of facial motion and their locations should be invariant against similarity transform (scale, rotation and translation). To this end, I propose to generate the feature points by randomly sampling around the local regions of reference 2D landmarks (the mean of all training images' 2D landmarks) and index them by the barycentric coordinates with respect to the closest Delaunay triangles formed by those landmarks as in (C. Cao et al., 2014). Before each stage regression, I first generate M feature points and save the corresponding triangle indices and barycentric coordinates. Then, $\mathcal{F}^t(\cdot)$ calculates 2D landmark positions from the previous facial shape estimation - $\{\alpha, \mathrm{Q}, P^{t-1}\}$ and calls the saved indexing information to extract pixels from the image.

3.5 Experiments

This section first validates the proposed GoMBF and GoMBF-Cascade regression in 3D facial tracking on in-the-wild videos. Then it deeply investigates the effect of various synthetic data on training GoMBF-Cascade for 3D facial tracking.

Implementation. The number of fern levels - $F + 1$ balances the fern's fitting power in training and its generalization ability in testing. I follow the most common setup in previous studies (X. Cao et al., 2014) and set F as 5, which generates promising results in my experiment. It can also be found that via increasing the number of stages – T, the number of ferns - $K_{\boldsymbol{\delta}}, K_{\boldsymbol{\theta}}, K_{\mathbf{t}}, K_{\mathbf{D}}$ and the number of feature points – M, the tracking accuracy increases, but the runtime speed slows down and the memory cost increases. For balancing between computational cost and accuracy, these parameters are empirically chosen as $T = 10$, $K_{\boldsymbol{\delta}} = K_{\boldsymbol{\theta}} = K_{\mathbf{t}} = K_{\mathbf{D}} = 80$ and $M = 600$. For the number of initial expressions - L in runtime prediction, I tested four different values – 10, 20, 30, 40 and found 20 produced a promising visual tracking result without introducing too much computational cost. Overall, the parameters of GoMBF-Cascade regression are set as follows: offline training - $T = 10$, $F = 5$, $K_{\boldsymbol{\delta}} = K_{\boldsymbol{\theta}} = K_{\mathbf{t}} = K_{\mathbf{D}} = 80$, $M = 600$; runtime prediction - $L = 20$. This parameter configuration applies to all the following experiments without further specification. Since this work focuses on accurate facial motion regression, there is no post-processing and parametric face model adaption during online tracking as in previous studies (C. Cao et al., 2014; Saito et al., 2016). The focal length f is empirically set as 1,000 and the facial identity coefficients $\boldsymbol{\alpha}$ are estimated from the first video frame and keep fixed for the rest frames. Whereas the setup is somewhat rough and poses much bigger challenges on facial motion regression, GoMBF-Cascade is able to produce accurate and temporally-smooth tracking results. The tracking system is implemented using C++ with OpenMP parallelization, and tested on a laptop with a quad-core Intel Core i5 (2.30GHz) CPU and an integrated web camera producing 640 x 480 video frames. The system achieves a 30fps performance.

3.5.1 GoMBF-Cascade Validation

1) Datasets

Training Data. Ideally, the proposed method requires an image dataset with accurate 2D landmark annotations and ground-truth 3D shape parameters that match the parametric face model for training. However, there is no such data available. As an alternative, I select images from three public face datasets and generate the corresponding 2D/3D labels by myself. The training images are from 300W-3D (X. Zhu et al., 2016), FaceWarehouse (C. Cao, Weng, Zhou, et al., 2013) and Multi-PIE (R. Gross et al., 2010):

300W-3D contains 3,837 in-the-wild face images, each being offered with 68 hand-labelled landmarks (I discard the two points on the inner mouth corners in this work) and a reconstructed 3D facial mesh. For each image, I first estimate identity and expression coefficients - $\{\boldsymbol{\alpha}, \boldsymbol{\delta}\}$ by fitting the parametric face model to the provided 3D facial mesh based on landmark constraints. The fitting process resembles the one expressed in Eq. 3.10 with the only difference that E_{lan} measures 3D landmark distances in world space this time. I then get $\mathbf{Q}$ by fixing the focal length f to 1,000 and estimate the rotation and translation parameters - $\{\boldsymbol{\theta}, \mathbf{t}\}$ using the POSIT algorithm (Dementhon & Davis, 1995). Finally, $\mathbf{D}$ can be easily calculated by comparing the 2D landmarks projected from the estimated 3D face to the hand-labelled landmarks.

FaceWarehouse consists of 3,000 near-frontal face images captured from 150 human subjects under controlled indoor environment. I choose 1,600 images of 80 subjects to use in my experiment and detect 66 landmarks for each image using a pre-trained SDM model (Xiong & De la Torre, 2013). Since the algorithmic landmark detection is not accurate enough, I go through all the images and manually adjust the misaligned landmarks. I later follow the process explained in Eq. 3.10 to estimate 3D shape parameters from 2D landmark labels. To correct implausible facial expression estimations, I further manually tune the expression coefficients. The identity and head pose parameters are updated afterwards to align

Table 3.1: Training and testing datasets

Training Set	
300W-3D	3,837 images, >500 subjects
FaceWarehouse	1,600 images, 80 subjects
Multi-PIE	1,024 images, 63 subjects
Testing Set	
300VW	004, 007, 009, 018, 019, 028, 037, 044, 048, 119, 143, 205, 208, 213, 223, 224, 405, 524, 531, 558.
Live Video Streams	

with the new facial expression coefficients using a similar fitting method as described above.

Multi-PIE provides more than 4K indoor face images captured from 337 subjects. The images cover various facial expressions, head poses and illumination conditions. Each image has been manually annotated with 68 landmarks. I select 1,024 images of 63 subjects for training the facial motion regression. The corresponding 3D facial data is obtained with the same approach used for processing the FaceWarehouse data.

Overall, I collected 6,461 images for training. Table 3.1 shows the basic information of the training set. Despite the relatively smaller size of training set, the proposed GoMBF-Cascade regression delivers tracking results competitive to the state-of-the-art method (Y. Guo, Zhang, Cai, Jiang, et al., 2018) that used much more training data.

Testing Data. The tracking system has been evaluated on 20 challenging in-the-wild videos from 300VW (Shen et al., 2015). Each video records the facial performance of a human subject in an unconstrained environment. The videos have been labelled with 68 2D landmarks frame by frame, providing a good benchmark to assess the proposed tracking system that also outputs 2D landmarks. After scrutinizing the videos, I discard those that cannot be tracked since the first frame and then randomly select 20 videos from the rest of 300VW. The corresponding

video information is listed in Table 3.1. In addition, the tracking system has also been tested on live video streams.

2) Comparison with State-of-the-art Methods

I first validate GoMBF by comparing GoMBF-Cascade with the Explicit Shape Regression (ESR) method (X. Cao et al., 2014). To the best of my knowledge, besides deep learning methods (Y. Guo, Zhang, Cai, Jiang, et al., 2018) which I have compared below, all the other state-of-the-art learning-based 3D facial tracking approaches (C. Cao et al., 2014; C. Cao, Weng, Lin, et al., 2013; McDonagh et al., 2016) use ESR to build the facial motion regression. What's more, GoMBF-Cascade is closely connected with ESR. The main difference between them is that GoMBF-Cascade employs GoMBF instead of the conventional boosted ferns as the stage regressor. These make ESR a natural and good baseline for evaluating GoMBF and GoMBF-Cascade. For fair comparison, I implement ESR with the same appearance vector extraction function $\mathcal{F}(\cdot)$ and other setups as that used in GoMBF-Cascade regression: training - $T = 10$, $F = 5$, $K = 320$, $M = 600$; runtime prediction - $L = 20$. ESR and GoMBF-Cascade are then trained on the same training set as introduced above. I test the two tracking models - GoMBF-Cascade and ESR on the selected 300VW videos. The tracking results are evaluated both quantitatively and visually.

For quantitative comparison, I apply the widely-accepted point-to-point root mean square error (normalized by the face's inter-ocular distance) between the tracked 2D landmarks and the ground-truth annotations (Shen et al., 2015). For each video, I report the error averaged over all landmarks and video frames. As shown in Fig. 3.4, GoMBF-Cascade delivers lower 2D landmark tracking error for most of the 300VW videos than ESR. The average tracking error across all 20 videos is 0.0410 for GoMBF-Cascade and 0.0434 for ESR. Comparing against ESR, GoMBF-Cascade's average tracking error declines about 5.53%. In addition to the improved runtime tracking performance, GoMBF-Cascade exhibits faster convergence than ESR during training (see Fig. 3.5, the error is measured as the

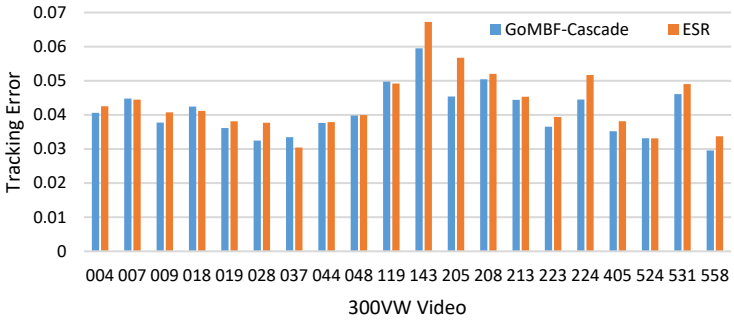


Figure 3.4: Comparison between GoMBF-Cascade and ESR on 2D landmark tracking.

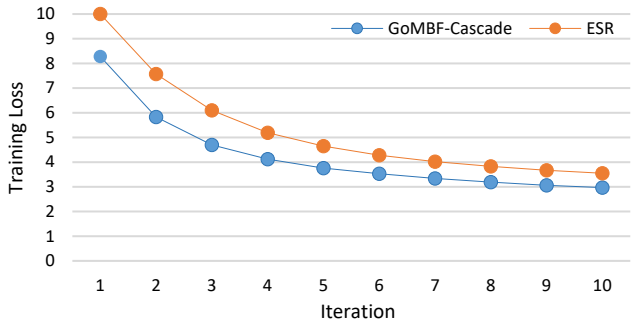


Figure 3.5: The training convergence curves of GoMBF-Cascade and ESR.

RMSE between the 2D landmark annotations and predictions). I believe that this benefits from the mechanism of GoMBF which exploits the complementary information between inter-modular boosted ferns and refines fern leaves towards the final regression target. What's more, GoMBF-Cascade enables parallel training when learning the modular boosted ferns. This significantly reduces the training time as compared to the traditional boosted ferns which has to be learned sequentially. In my experiment, it took about 3,532s to train a stage boosted ferns in ESR, while it only took 1,228s for GoMBF-Cascade, saving about 65.2% training time. I also visually compare GoMBF-Cascade and ESR by rendering out the tracked 3D faces. As shown in Fig. 3.6, GoMBF-Cascade is able to track facial expressions especially the mouth movements more precisely than ESR. GoMBF-Cascade is also found to be much more resilient to occlusions than ESR (see Fig. 3.7). In conclusion, both quantitative and visual results demonstrate that the

Figure 3.6: GoMBF-Cascade tracks facial expressions more accurately than ESR.

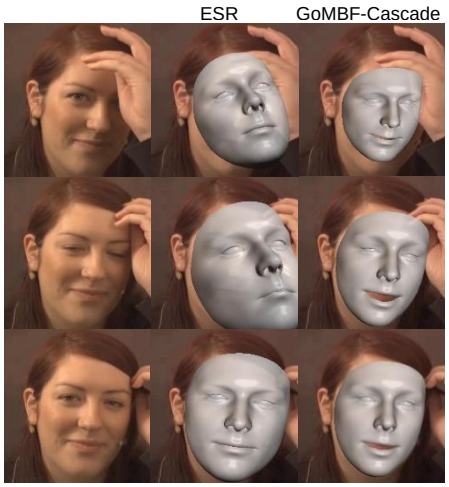


Figure 3.7: GoMBF-Cascade shows higher resilience to occlusions than ESR.

proposed GoMBF outperforms the conventional boosted ferns in fitting power and learning speed.

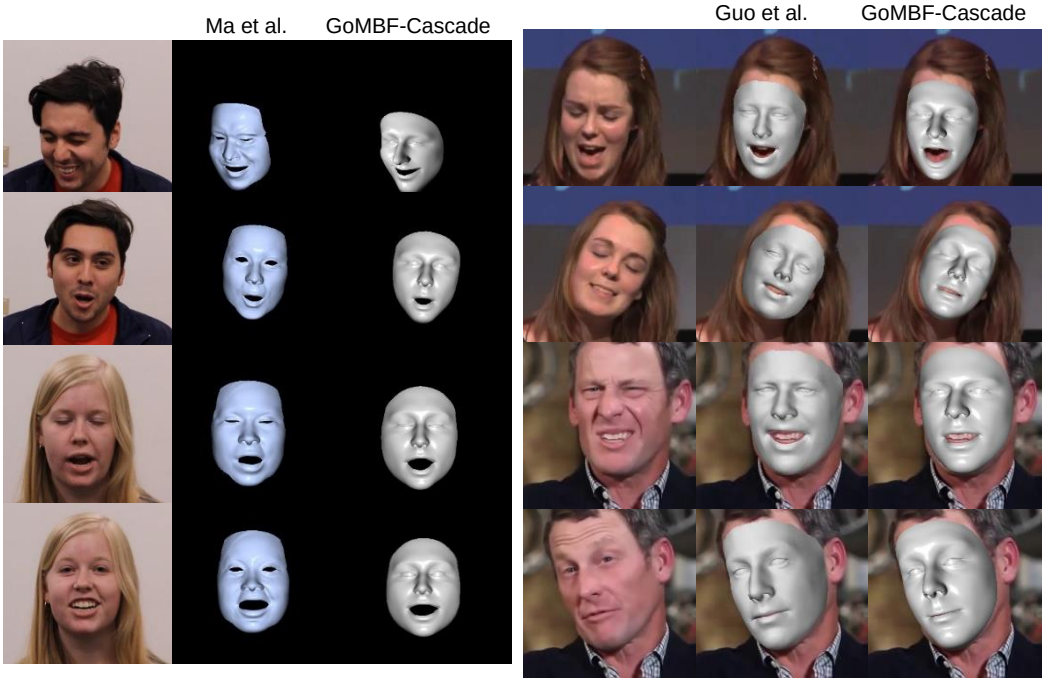


Figure 3.8: Comparison between GoMBF-Cascade and Ma et al. (Ma & Deng, 2019) and Guo et al. (Y. Guo, Zhang, Cai, Jiang, et al., 2018). Please note that the authors of (Ma & Deng, 2019) provided me the videos for testing and comparison.

To further verify the robustness of GoMBF-Cascade regression, I compare its tracking results with those output from two state-of-the-art 3D facial tracking approaches - (Ma & Deng, 2019) and (Y. Guo, Zhang, Cai, Jiang, et al., 2018). (Ma & Deng, 2019) is a typical optimization-based method which casts the tracking process into minimizing a highly non-linear objective function that enforces alignment on sparse feature points and pixel intensities between the reconstructed 3D face and the input video frame. It relies on GPU computing to achieve real-time performance. (Y. Guo, Zhang, Cai, Jiang, et al., 2018) instead resorts to convolutional neural networks (CNN) to regress facial shape and appearance parameters from facial images. It used 80K facial images to train its tracking network. Please note that the authors of the aforementioned two approaches only provided me the visual tracking results of their approaches. I thus cannot quantitatively compare those two approaches with GoMBF-Cascade using the landmark-based metric. Alternatively, I report the visual comparison results in

Figure 3.9: Tracking results of GoMBF-Cascade on live video streams.

Fig. 3.8. It is worth pointing out that, due to its intuitive nature, visual comparison has also been frequently used for evaluating the method's performance in the field of 3D facial tracking. As shown in Fig. 3.8, GoMBF-Cascade achieves competitive tracking performance against the two methods that either relied on an intricate photo-geometric fitting process (Ma & Deng, 2019) or was trained on a large-scale dataset (Y. Guo, Zhang, Cai, Jiang, et al., 2018). It does better on tracking eye closure and mouth movements than those two methods. Furthermore, tracking results on live video streams also demonstrate the robustness of GoMBF-Cascade (see Fig. 3.9). Please note that my tracking results are purely based on the proposed GoMBF-Cascade regression without any post-processing on the regressed expression and head pose parameters. As demonstrated, GoMBF-Cascade provides a robust and elegant solution to 3D facial tracking with a reasonably small set of training data. (For more tracking results, please refer to the supplementary video)

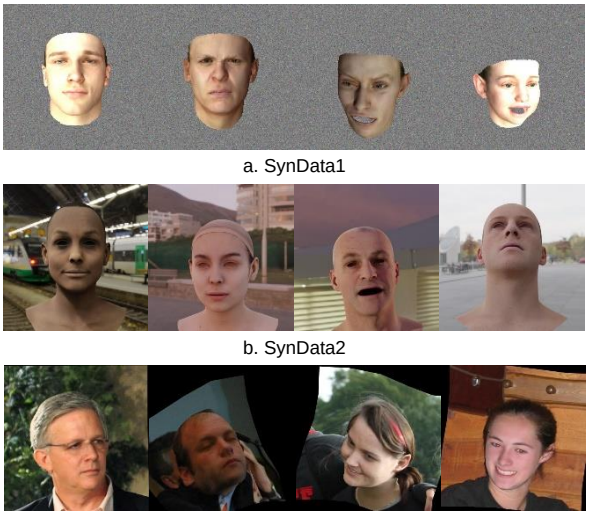
3.5.2 Training with Synthetic Data

As described above, collecting facial images with accurate 3D geometry is tedious, which normally needs time-consuming human inspection and correction. As an alternative, synthesizing facial imagery for training is highly-efficient and provides fully accurate 3D labels. Whereas this novel data harvesting method has been successfully applied in 3D facial tracking and reconstruction using deep learning (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2017, 2016), it remains unclear if the synthetic data also favours non-deep learning methods such as GoMBF-Cascade. This part investigates the largely unexplored problem by using three types of synthetic facial imagery with various naturalness levels to train GoMBF-Cascade. The tracking models trained on real data, on synthetic data and on a mixture of data are then compared with each other.

1) Synthesizing Training Imagery

In computer graphics, simulating real-world lighting and facial texture is crucial in rendering photo-realistic faces. Based on this insight, I apply three different lighting and texture models to synthesize facial imageries with various naturalness levels:

At the first stage, I incorporate BFM's texture components (Paysan et al., 2009) and Phong illumination (Phong, 1975) into the parametric face model to render new faces. To cover a wide range of facial shapes, poses and lighting conditions, I construct multiple groups of rendering parameters. Specifically, I generate 40 3D heads by randomly sampling shape and texture coefficients from the corresponding normal distributions provided by BFM. For each head, 30 samples in various poses are generated, including 10 with neutral expression and specified head poses, 10 in frontal pose but with specified expressions, and 10 with random head poses and expressions (head pose and expression coefficients are chosen from the pre-built 300W-3D dataset). To render each head sample, I randomly select four lighting conditions from a set consisting of 72 Phong illumination models



a. SynData1

b. SynData2

c. SynData3

Figure 3.10: Synthesized facial images.

which vary in components of direction, specular reflection, diffuse reflection, ambient reflection and shininess. In total, I synthesize 4,800 3D head samples. After defining a perfect pinhole camera model with a focal length of 1,000 and setting the image size to 450 x 450, all the 3D heads are rendered to images with a background filled with Gaussian noise (see Fig. 3.10a for examples).

As BFM (Paysan et al., 2009) and Phong illumination (Phong, 1975) only simulate the facial texture and scene lighting in a coarse level, the synthetic faces from the first stage look rough and present clear artefacts. To improve the naturalness of the synthesized faces, I utilise a bundle of 20 high-quality head texture maps captured with a commercial photogrammetry rig[1]. The texture maps pair with two base meshes of an identical topology. Each texture occupies exactly the same UV and can be swapped out conveniently for a different texture, hence resulting in 40 different 3D heads. Since the base mesh is in repose, I generate its delta blendshapes using deformation transfer (Sumner & Popović, 2004) to enable facial expression modelling. For each base mesh, I also manually annotate 66 landmarks which share the same semantic meaning as those used in my parametric

[1] https://www.3dscanstore.com/.

Figure 3.11: Materials for generating SynData2. Top row: high-quality texture map and textured 3D facial mesh; middle row: HDR images; bottom row: the synthesized facial images.

face model. With the matched 3D landmarks, I can easily estimate from the synthesized 3D head the required shape parameters using the approach mentioned in 300W-3D data processing. To realistically illuminate the head, I apply an image-based lighting technique in which high dynamic range (HDR) panoramic images are used to provide the environment lighting. The technique captures omni-directional light information of a real-world scene and stores it into pixels of a HDR image which can be projected to a sphere simulating the surrounding space of the target object. I collect 12 HDR images (see Fig. 3.11) which were captured from common indoor and outdoor scenes such as train station, hotel room and misty pines. Following the procedure as described in the first stage, I generate 30 samples in various poses for each 3D head. Each head sample is then rendered to 450 x 450 images with four lighting conditions and the image's background is set as the scene exhibited in the corresponding HDR image for more natural synthesis. I use the inbuilt Cycles path-tracing engine of Blender[2] for rendering. As shown in Fig. 3.10b, highly photo-realistic facial images with fine texture features such as pores and wrinkles can be synthesized.

[2] https://www.blender.org/.

Table 3.2: Synthetic dataset

SynData1 (4,800 images)	- 40 3D heads randomly generated from BFM. - Each head is rendered with 30 different poses and 4 Phong illumination conditions.
SynData2 (4,800 images)	- 40 3D heads with very high-quality texture. - Each head is rendered with 30 different poses and 4 natural lighting conditions simulated with HDR images.
SynData3 (9,300 images)	- Selected from CoarseData which was built by applying lighting and texture estimated from in-the-wild facial images.

Comparing to in-the-wild data, facial images synthesized in the first two stages still have pronounced artefacts, e.g. the lack of inner-mouth structure, limited variations in facial shape, lighting and background. In this stage, I turn to another kind of synthetic data (Y. Guo, Zhang, Cai, Jiang, et al., 2018) which is derived from using facial shape, texture and lighting estimated from real-world images for rendering. By further warping the background region of the source image to fit the new face, the synthetic image can look very similar to real-world counterpart. Guo et al. (Y. Guo, Zhang, Cai, Jiang, et al., 2018) have released such a dataset which was named CoarseData (see Fig. 3.10c). The dataset was generated from 3,131 300W-3D images, with each image being augmented 30 times to cover more facial expressions and head poses. In my experiment, I randomly select 9,300 images (about 3 samples for each original 300W-3D image) from CoarseData and apply the same method used in processing the 300W-3D data to get the ground-truth shape parameters that fit the parametric face model.

For convenience, I call the training set built in Section 3.5.1 as RealData, the three synthetic datasets as SynData1, SynData2 and SynData3 respectively. The corresponding information is listed in Table 3.2.

2) Tracking Model Comparison

Taking the GoMBF-Cascade regression trained on RealData as the baseline model, I evaluate the models trained purely on synthetic data or on a mixture of real and synthetic data in tracking 300VW videos.

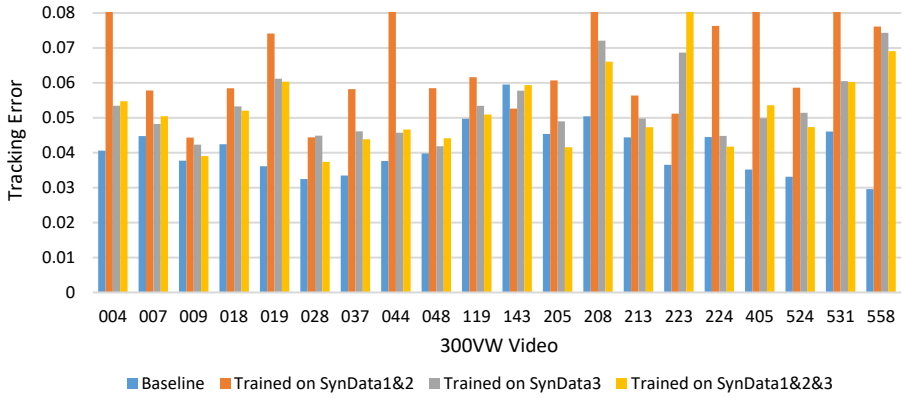


Figure 3.12: Comparison between the tracking models trained purely on synthetic data and the baseline model (error $\gg$ 0.08 has been cut off).

Training on synthetic data. To introduce sufficient facial shape and appearance variations during regression learning, I train three GoMBF-Cascade models which are on the mixture of SynData1 and SynData2 (SynData1&2), on SynData3, and on all the synthesized facial images (SynData1&2&3) respectively. I follow the same setups as training with RealData. The three models are then tested on 300VW videos and compared to the baseline model by calculating the aforementioned 2D landmark tracking error. As shown in Fig. 3.12, the models trained purely with synthetic facial images output much bigger tracking errors than the baseline model, especially the model trained with SynData1&2 which completely lost the face in some videos such as video-044 (with an error of 110.47). Even for the models trained with SynData3 which comprises synthesized facial images looking very close to the real in-the-wild data, its tracking accuracy is still sharply lower than the baseline model's accuracy. For those four tracking models, the average tracking error across all 300VW videos is 0.0410 (Baseline), 0.0525 (SynData1&2&3), 0.0534 (SynData3) and 5.6954 (SynData1&2) respectively. Comparing with the baseline model, the models trained with synthetic data show different degrees of increase in tracking error: 28.05% (SynData1&2&3), 30.24% (SynData3), 13791.22% (SynData1&2). Their corresponding standard deviations - 0.0115 (SynData1&2&3), 0.0096 (SynData3) and 24.6656 (SynData1&2) are

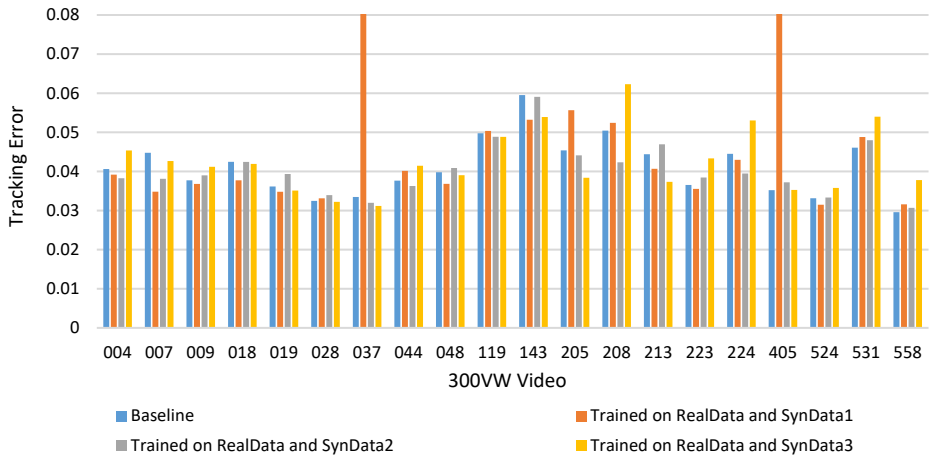


Figure 3.13: Comparison between the tracking models trained on the mixture of data and the baseline model (error $\gg$ 0.08 has been cut off).

also much higher than the baseline model's one - 0.0073 (Baseline). These metrics further demonstrate that only using synthetic data for training does not introduce an overall improvement in tracking performance, but instead it makes the tracking model less stable.

It is worth pointing out that McDonagh et al. (McDonagh et al., 2016) successfully learned an ESR-based facial motion regression for personalized 3D facial tracking from synthesized training imageries. However, their synthetic images were rendered from a high-quality facial rig (built from an offline facial capture system) that fits tightly with the user's facial geometry and appearance, and an illumination model driven by light probe data acquired at the target environment. This can hardly be achieved in unconstrained facial tracking scenario where the target environment and user are unknown in the training phase.

Training on mixed data. To further investigate the impact of synthesized facial images, I sequentially mix the synthetic data with the real data for training GoMBF-Cascade. As a result, I generate three tracking models. Fig. 3.13 presents the 2D landmark tracking errors of these three models on 300VW videos. As shown in the figure, for most videos, at least one of the three models trained on the mixed data exhibits improved tracking performance than the baseline model.

However, none of them is as reliable as the baseline model that can be generalized well to all the testing videos. The model trained with SynData1 even outputs an extremely large tracking error – 19.56 on video-037. The statistics of the average tracking error across all videos also support this observation. Compared with the baseline model, although there is a slight decline of 1.46% in average tracking error for the model trained with SynData2, much bigger increases of 2401.95% and 3.66% are observed for the models trained with SynData1 and SynData3.

From these two experiments, I find that: i) the GoMBF-Cascade tracking model trained purely on synthesized facial images cannot generalize well to unconstrained real-world data; ii) involving synthetic facial images into training benefits tracking in some certain scenarios, but degrades the tracking model's generalization ability. Interestingly, these two findings are contrary to those observed in deep learning-based 2D/3D facial tracking and reconstruction (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Kortylewski et al., 2018; Richardson et al., 2017, 2016). As reported in (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2017, 2016), facial tracking/reconstruction CNN models that work well on real-world data can be learned only with similar synthetic facial images as SynData1 and SynData3. In (Kortylewski et al., 2018), the authors find that priming deep networks by pre-training them with synthetic facial images is helpful for reducing the negative effects of the training data bias. Presumably, the discrepancy is mainly caused by the different feature learning capabilities between deep and non-deep learning methods.

3.6 Conclusion

This chapter first developed a novel regression method called GoMBF-Cascade for real-time 3D facial tracking from a monocular RGB video. GoMBF-Cascade is mainly featured with a sequence of globally-optimized modular boosted ferns – GoMBF, which is built with compositional learning and can efficiently handle the modality variety in facial motion parameters during regression. Compared with the

conventional boosted ferns (C. Cao et al., 2014; C. Cao, Weng, Lin, et al., 2013; X. Cao et al., 2014), GoMBF exhibits stronger fitting power and a higher learning speed. In theory, GoMBF can be seamlessly adapted to any other multi-output regression tasks. The resulting GoMBF-Cascade regression has been validated in 3D facial tracking on in-the-wild videos and live video streams. It delivered competitive tracking performance comparing against the state-of-the-art methods (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Ma & Deng, 2019) which require a large-scale training set or have a much higher computational complexity, hence providing a robust and highly elegant solution to real-time 3D facial tracking.

The chapter also systematically investigated the effect of synthesized facial imageries on training GoMBF-Cascade for 3D facial tracking. It applied three different kinds of synthetic facial images with various naturalness level for training and compared the tracking models trained on real data, on synthetic data and on a mixture of data. The experimental results showed that, i) training purely with synthesized facial images can hardly deliver a robust 3D facial tracking model that generalizes well to unconstrained real-world data; ii) involving synthetic images into training can benefit tracking in some certain scenarios, but harms the tracking model's generalization ability. This provides a different understanding of learning from synthetic facial images as those formed in deep learning-based 2D/3D facial tracking and reconstruction (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Kortylewski et al., 2018; Richardson et al., 2017, 2016). It is supposed to be caused by the different feature learning capabilities between deep and non-deep learning approaches. I believe the findings can benefit a series of non-deep learning facial image analysis tasks where the labelled real data is difficult to access.

It can be found that, by conditioning on facial pose, expression and illumination, the generative adversarial network (GAN) (Tewari et al., 2020) is able to synthesize extremely realistic facial images. This provides a highly flexible and efficient way for synthesizing facial imagery with ground-truth labels. In the future, it would be a very promising direction to apply such networks to generate the training data for 3D facial tracking.

Chapter 4
Realistic 3D Facial Expression Reconstruction for VR HMD Users

Foreword

With the advent of virtual reality (VR) technology, humans are able to interact with a simulated 3D world close to or completely different from the real world in a seemingly real or physical way. Wearing a head-mounted display (HMD), the user can communicate with each other or interact with the virtual world with full immersion, an experience that cannot be afforded by traditional video communication. However, the VR HMD significantly occludes the user's face, making it intractable to capture the user's facial performance with conventional vision-based methods, such as those developed in the previous two chapters. To solve this problem, this chapter proposes to combine a robust monocular 3D face reconstruction algorithm with an EMG-based facial biosensing technique – Faceteq (Mavridou et al., 2017) to achieve realistic 3D face capture of the VR HMD user. The proposed method offers a practical solution to enable face-to-face communication with compelling facial expressions and emotions in the VR context.

The chapter is based on a published journal paper:

- **Lou, J.**, Wang, Y., Nduka, C., Hamedi, M., Mavridou, I., Wang, F. Y., & Yu, H. (2019). Realistic facial expression reconstruction for VR HMD users. *IEEE Transactions on Multimedia*, 22(3), 730–743.

In the paper, I designed and implemented most parts of the reconstruction system, including 3D face embodiment generation, FACS AUs prediction from EMG signals and emotion recognition from AUs. I also conducted the experimental validation and wrote up the paper.

4.1 Introduction

Recent progress in virtual reality (VR) has introduced immersive user experience in virtual worlds. Existing mainstream head-mounted displays (HMDs), such as Oculus Rift and HTC Vive enable users to perceive the virtual world, but they only allow limited interactions between the user and the virtual environment. These interactions are mainly based on human body motion capture and hand tracking technologies but ignore the importance of facial expressions for communication.

Facial expressions serve as the primary nonverbal means of communication among human beings (Ekman & Rosenberg, 1997). A truly interactive and immersive experience cannot be envisioned without the technologies for sensing and recovering the user's facial expressions in VR. However, VR HMDs usually occlude a large part of the user's face, which rules out most existing facial performance sensing methods, such as ordinary camera-based technologies. A few recent works (Cha et al., 2016; H. Li et al., 2015; Olszewski et al., 2016; Suzuki et al., 2017; Thies, Zollhöfer, et al., 2016) have explored solutions to this problem. Li et al. (H. Li et al., 2015) and Olszewski et al. (Olszewski et al., 2016) made preliminary trials of equipping HMDs with the facial performance capture ability. However, their solutions require a RGB-D or RGB camera mounted on the HMD, which are not ergonomically comfortable and cause an extra head burden. Some works resorted to other advanced facial sensing technologies, such as infrared (IR) sensors (Cha et al., 2016) and (Suzuki et al., 2017) electromyography (EMG) sensors (Gruebler & Suzuki, 2014; Mavridou et al., 2017). These lightweight optical or contact-based sensors can be easily embedded into the headset in an unobtrusive manner, thus open a new era of HMD-based wearable facial performance sensing systems.

However, existing solutions (H. Li et al., 2015; Olszewski et al., 2016; Suzuki et al., 2017) build the HMD user's face embodiment with non-realistic facial shape or texture. Moreover, a few systems (Suzuki et al., 2017) can only detect the facial

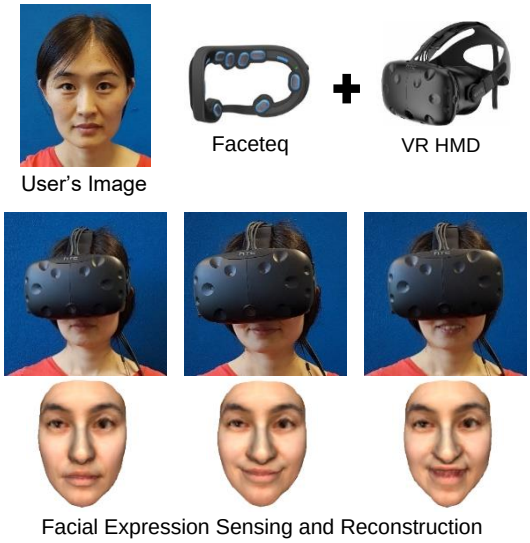


Figure 4.1: A demonstration of the developed system. With a single face image, it is able to sense and reconstruct realistic facial expressions of the head-mounted display (HMD) user. Faceteq is a facial sensing wearable device that can be attached on mainstream HMDs. By utilizing eight integrated electromyography (EMG) sensors, Faceteq enables the detection of the facial muscle contractions of the HMD wearer.

expression category which is subsequently represented with pre-defined facial movements on a pre-made virtual avatar. This prohibits natural interactions between participants in the virtual world. To address these problems, this chapter develops a framework (see Fig. 4.1) that captures facial activities coded in facial action units (AUs, see Fig. 4.2) (P. & Friesen, 1978) upon an advanced facial sensing hardware and can exhibit realistic expressions via a compelling digital embodiment of the user's face in virtual scenarios.

The proposed system embeds a pioneering facial sensing hardware – Faceteq^{TM} (Mavridou et al., 2017) into the HMD to detect facial muscle activities through integrated EMG sensors placed on the most emotionally salient facial part (ESFP) – the eye region. Relevant AU-coded facial expressions are then identified with a machine learning method from pre-processed Root Mean Square (RMS) levels of recorded EMG signals. With a single image of the user, the system reconstructs the user's 3D face and generates AU-based blendshapes using a

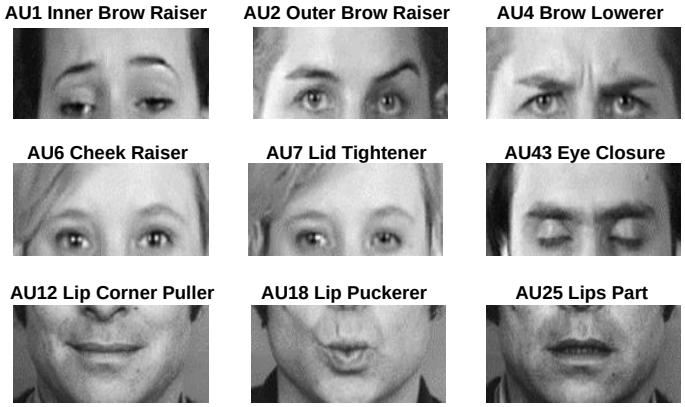


Figure 4.2: Examples of facial action units (Martinez et al., 2017).

popular analysis-through-synthesis approach (Thies, Zollhofer, et al., 2016) and a robust non-rigid shape registration algorithm (S. Zhang et al., 2017). Classic psychological studies predict basic emotions from AUs following a few heuristic rules (P. & Friesen, 1978). However, as each category of emotions can have slightly different muscle group contraction, the real relationship between AUs and emotions turns to be complicated, which can be hardly explained with limited rules. The approach becomes infeasible when the observed AUs are not covered by the rules, such as in my case where seven AUs (AU1, AU2, AU4, AU6, AU12, AU25 and AU43) focusing on the ESFPs – the eye and mouth region are investigated. To this end, the system uses the fern classifier to model the probability of emotions given activated AU information.

The system developed in this chapter has been validated through appropriate experiments. Here is a summary of the main contributions:

- It proposes the first automatic system of its kind that senses and reconstructs the HMD user's facial expressions with a realistic face embodiment.
- It develops an innovative correlation from facial biometric (EMG) signals to facial expressions through individual AUs, which explicitly captures the detailed facial movements performed by the HMD user.

- It proposes a novel probabilistic model that builds relationships between AUs and the six basic emotions.

4.2 Related Work

4.2.1 HMD-based Facial Sensing Systems

The VR HMD occludes a significant portion of the user's face, preventing most existing facial performance capture approaches from being applied. Some recent works embedded small advanced optical sensors inside the headset, such as IR sensors (Cha et al., 2016; Suzuki et al., 2017) to recognize facial expressions of basic emotions by detecting facial movements. Meanwhile, the contact-based sensing technology is drawing great attention in facial wearable device application because of its superiority in scenarios with highly constrained visibility. The electroencephalography (EEG) has been used in (Wolpaw et al., 2002) to record brain activities to detect basic emotions, but extensive training and user concentration are required. A commercial device from Looxid Labs[3] incorporates two-channel EEG into the HMD, but whilst this may provide some information of user focus, this will not provide information that directly relates to user valence and facial expressions. An alternative is to measure the electrical signals of muscle activities. EMG is more sensitive at detecting micro-contractions of muscles and indeed was used to calibrate initial computer vision facial tracking algorithms (Cohn & Schmidt, 2003). It has been successfully combined with facial wearable devices for recognizing facial expressions and emotional states (Gruebler & Suzuki, 2014; Mavridou et al., 2017). All these technologies offer a wide range of pathways to a reliable HMD-based facial sensing system. However, the studies above only predict from biometric signals facial expression categories (e.g. happiness, anger) whereby what facial movements were involved is still unknown.

[3] http://looxidlabs.com/

Li et al. (H. Li et al., 2015) first equipped the HMD with the ability to sense almost the whole face region. They integrated the HMD with eight strain gauges and an RGB-D camera to capture facial activities of the occluded upper face region and the mouth. The captured facial performance data were then mapped to blendshape coefficients through a linear regression to realize real-time facial animation. However, their solution requires tedious calibrations for each user and the mounted RGB-D camera introduces an extra head burden. Olszewski et al. (Olszewski et al., 2016) subsequently improved Li's solution with a lightweight RGB camera and two IR cameras that see direct views of the HMD user's mouth and eyes, and the convolutional neural network that builds a robust mapping between facial images and animation parameters of a pre-built virtual character. Their approach is free of user-specific calibrations and can generate relatively comprehensive facial animation.

Although existing techniques and approaches have pushed the boundary of HMD-based facial sensing systems forward, the following three problems remain unresolved:

a) Most previous systems (Cha et al., 2016; Gruebler & Suzuki, 2014; Mavridou et al., 2017; Suzuki et al., 2017) concentrate on recognition of facial expression categories while ignoring the fact that facial expression has multiple appearance representations. For example, happiness can be expressed with either the AU12 (lip corner puller) or a combination of AU12 and AU6 (cheek raise), so a deeper insight into the composition (e.g. AUs) of facial expressions is needed.

b) The fidelity of reconstructed facial expressions has not attracted sufficient attentions yet. Previous works (H. Li et al., 2015; Olszewski et al., 2016) only consider the geometry of the facial expression while omitting the 3D shape and texture of the user's face, which would generate unrealistic facial expressions in virtual environment.

c) Previous studies have not created an integrated pipeline whereby a personalized model of the user's face can be captured with a smartphone and used to better represent how they express themselves.

4.2.2 3D Face Reconstruction from a Single Image

The field of 3D face reconstruction from a single image has witnessed significant progresses over the past two decades (Blanz & Vetter, 1999; Romdhani & Vetter, 2005; Sela et al., 2017; Sengupta et al., 2018; Thies, Zollhofer, et al., 2016). Readers are referred to (Zollhöfer et al., 2018) for a comprehensive survey. Researches differ along two main dimensions, the underlying face prior and the reconstruction algorithm.

1) Face Priors

Face priors that model the geometry and texture of faces typically serve as the basis of 3D face reconstruction in the ill-posed monocular setting. 3D Morphable Model (3DMM) is a statistical face prior which was originally generated from a database of 200 scanned neutral human faces (Blanz & Vetter, 1999). It depicts the facial geometry and albedo within a multi-linear Principal Component Analysis (PCA) subspace. The 3DMM can also be extended to faces with expressions (Jiang et al., 2018; X. Zhu et al., 2016). Although 3DMM has laid the foundation for monocular 3D face reconstruction, it shows limitations when dealing with in-the-wild data captured in uncontrolled scenarios. With the available large-scale scanned 3D/4D face data (Booth et al., 2018; S. Cheng et al., 2018; T. Li et al., 2017) and in-the-wild texture modelling (Booth et al., 2017), 3DMM has been pushed closer to fully solve the 3D face reconstruction from unconstrained images. Apart from 3DMM, a single 3D face reference has also been applied (Kemelmacher-Shlizerman & Basri, 2010). As this reference model should provide an initial estimation of the facial geometry and albedo, it thus needs to closely depict the desired face.

2) Reconstruction Algorithms

There are two main lines in this phase: generative approaches (Aldrian & Smith, 2012; Blanz & Vetter, 1999; Thies, Zollhofer, et al., 2016) and discriminative approaches (Sela et al., 2017; X. Zhu et al., 2016). The generative approach treats

the monocular 3D face reconstruction as an inverse rendering problem and formulates it as a complex optimization process. Metrics such as the color consistency, feature similarity and regularization constraints are then used to direct the optimization (Romdhani & Vetter, 2005; Thies, Zollhofer, et al., 2016). This kind of approach can recover promising 3D facial geometry and texture information, while achieving real-time performance with the GPU solver (Thies, Zollhofer, et al., 2016). However, the inverse rendering problem is highly ill-posed due to the incomplete input data, it is hence prone to degenerating in challenging scenarios, such as dealing with faces under severe occlusions and large poses.

Recently, the discriminative approach has emerged as an essential research branch resulting from dramatic progress in deep learning (Sela et al., 2017; Sengupta et al., 2018). Built on top of a database that contains extensive face images as well as the corresponding 3D facial data, deep learning is able to embed massive image-face relationships into a robust non-linear regression. This alleviates the problem of ill-posed images which creates incomplete or uncontrolled input data. Existing 3D face databases (C. Cao, Weng, Zhou, et al., 2013; Yin et al., 2006) were collected with sophisticated 3D facial capture systems in controlled settings or using the aforementioned generative approaches on in-the-wild face images (X. Zhu et al., 2016). Such processes are time-consuming, labor-intensive or not able to provide the fine-scale 3D facial data. To tackle these problems, a few recent studies resorted to the synthetic 3D facial data and yielded impressive results (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2016). This is in line with an interesting theory of parallel vision (K. Wang et al., 2017) which discusses the significance of synthetic data in addressing the problems of visual perception and understanding.

The combination of the generative approach and the discriminative approach now appears as an important direction in this (Tewari et al., 2018, 2017). The state-of-the-art (Tewari et al., 2018) that integrates a convolutional encoder network with an expert-designed generative model does not require any 3D facial data for training, while is still able to output promising reconstructions. Recovery of facial

geometry and texture details using deep neural networks is also an interesting direction (Huynh et al., 2018; Saito et al., 2017).

To eliminate the need of massive training data, this chapter turns to a practical and reliable generative approach (Thies, Zollhofer, et al., 2016) which has been validated in several state-of-the-art works (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Hu et al., 2017; Saito et al., 2017). It moves a step further to generate personalized blendshapes with a robust non-rigid registration method (S. Zhang et al., 2017) and a direct deformation transfer from generic AU-based blendshapes (H. Yu et al., 2012).

4.2.3 Emotions from Facial Action Units

The Facial Action Coding System (FACS) (P. & Friesen, 1978) is the best-known taxonomy of human facial expressions. It uses AUs to code the visually observable actions of individual or a group of facial muscles. For example, AU12 describes the contraction of the Zygomaticus major muscle that is typically observed in the expression of happiness. Thus, FACS AUs can offer a detailed interpretation of facial expressions resulting from facial muscle contractions. The revised FACS (Ekman, 2002) defines 32 anatomic AUs and 14 Action Descriptors (ADs) with respect to the head pose, gaze and other actions such as blow and bite. This chapter equips an EMG-based facial sensing hardware (Mavridou et al., 2017) with a learning method to identify common facial expressions coded in AUs.

Psychological studies (P. & Friesen, 1978) suggest that emotions such as happiness, sadness, fear, anger, surprise, and disgust can be predicted from AUs with a few heuristic rules (e.g. AU6, 12 normally indicates happiness). However, the rules cannot fully explain the complicated relationship between AUs and emotions, since each category of emotions can have various facial appearance representation. Existing rule-based methods (Valstar & Pantic, 2006; Velusamy et al., 2011) are thus unable to provide emotional information from AU combinations that are not included in the established rules. To address this problem, this chapter proposes to use the fern classifier (Ozuysal et al., 2007) to build the relationship

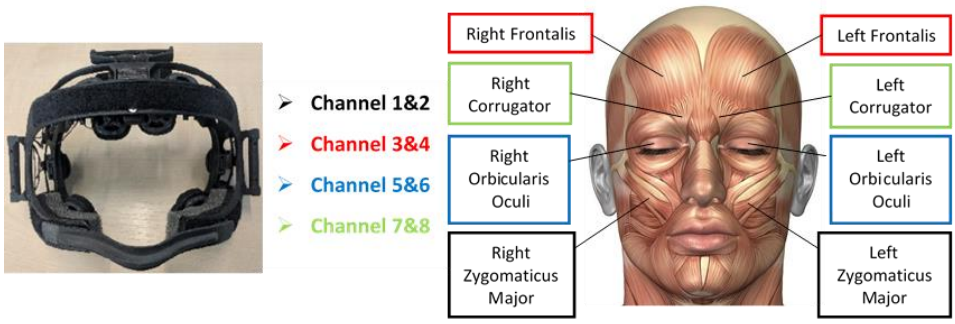


Figure 4.3: A prototype of Faceteq. The device equips with eight dry EMG sensors placed on the ESFP that output eight-channel EMG signals. Each channel correlates to a specific facial muscle (highlighted with rectangles in different color) contraction (the facial muscle picture is retrieved from https://fineartamerica.com/).

between six basic emotions and AUs with a posterior probability model. The proposed method fits well with the scenario where only AUs around the eye and mouth region are studied.

4.3 System Overview

4.3.1 Device

The developed system integrates a wearable facial sensing hardware - Faceteq (Mavridou et al., 2017) (see Fig. 4.3) that fits with mainstream commercial VR head-mounted displays (HMDs). The hardware utilizes the EMG technology to detect facial muscle activations. It consists of eight surface dry EMG sensors placed on the ESFP that do not require skin preparation, conductive gel and adhesive pads, while having a 24-bit signal resolution, 1kHz sampling rate, 20-450Hz signal bandwidth and no inter-sensor latency. Each EMG sensor accounts for a unique muscle action on Zygomaticus major, Frontalis, Orbicularis oculi and Corrugator. The output gives eight channels of muscle activations as well as their intensity scores at 1kHz when wired to a PC based VR system such as Oculus Rift or HTC Vive, or 25Hz via Bluetooth when using a smartphone. The learning

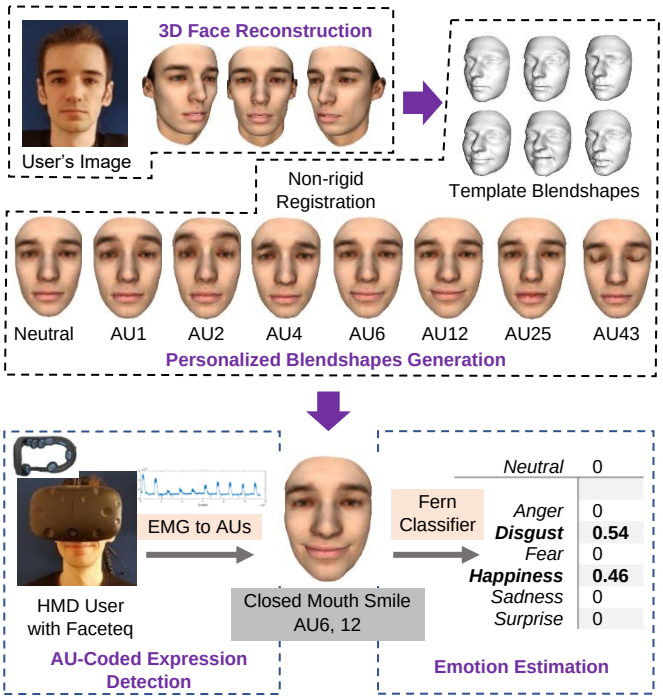


Figure 4.4: An overview of the proposed system. It first reconstructs a fully textured 3D face from a user's image. Then, template AU-based blendshapes are conformed to the reconstructed face to generate personalized blendshapes. With Faceteq and a robust learning method, the system predicts from EMG signals AU-coded facial expressions which are further reconstructed by fusing personalized blendshapes. A fern classifier is learned for estimating emotions from AUs.

method proposed in this chapter can further associate EMG channel activations to common AU-coded facial expressions.

The hardware contains two photoplethysmogram (PPG) sensors and an inertial measurement unit (IMU) including accelerometer and gyroscope which provide nine channels values of head movement, posture and state analysis at 50Hz. Meanwhile, it supports real-time signal quality monitoring, ASCII data files as well as binary files for post-acquisition data analysis.

4.3.2 Work Pipeline

As illustrated in Fig. 4.4, the system consists of four main parts: offline 3D face reconstruction, personalized blendshapes generation, online AU-coded facial expression detection and six basic emotions estimation.

To build a realistic digital embodiment of the user's face, the system only requires one photo image of the user. Then a state-of-the-art 3D face reconstruction approach (Thies, Zollhofer, et al., 2016) is applied to recover the dense 3D geometry as well as the texture of the user's face. The algorithm solves the reconstruction within a non-convex optimization process which takes photo consistency, sparse feature similarity and statistical regularization as constraints. The reconstruction is built upon a PCA-based morphable model (Paysan et al., 2009). FACS-based blendshapes (C. Cao, Weng, Zhou, et al., 2013; X. Zhu et al., 2016) and an illumination model based on Spherical Harmonics (Basri & Jacobs, 2003).

A template neutral facial mesh is warped to fit the reconstructed 3D face using a robust non-rigid registration method (S. Zhang et al., 2017). A linear rigid-alignment based on 68 facial landmarks is first used to estimate the pose between the template and the reconstruction. Then a coupled global and local deformation is applied to each point on the template to conform it to the reconstruction. Personalized blendshapes are obtained by simply transferring the deformations from a series of generic AU-based blendshapes (H. Yu et al., 2012) to the deformed template neutral face as they share the same topology. During the non-rigid registration process, the facial texture is transferred simultaneously with the geometric deformations based on the built point correspondences.

Eight-channel EMG signals from the facial sensing hardware are then fed into a learning method to predict common AU-coded facial expressions which drive personalized blendshapes to generate realistic facial expressions during online tracking. The learning method applies the Least-square Support Vector Machine (LS-SVM) on RMS features from EMG signals. Training and testing are conducted on a database collected from 15 confirmed mentally and physically healthy participants without any signs of conditions that could affect their face and thus facial expressions.

To get a deeper insight into the VR HMD user's internal emotional states, a probabilistic model is built to map AUs to six basic emotions. A fern classifier is

used to model the posterior probability of basic emotions given combinations of AUs. The fern classifier is learned from CK+ (Lucey et al., 2010) and EmotioNet (Fabian Benitez-Quiroz et al., 2016).

4.4 Face Embodiment Construction

With a single image, the developed system builds a digital embodiment of the user's face that contains a series of fully-textured 3D facial meshes in neutral pose or with AUs.

4.4.1 3D Face Reconstruction

Following Thies et al. (Thies, Zollhofer, et al., 2016), the system converts the 3D face reconstruction from a single image into an inverse-rendering problem which is solved through an analysis-by-synthesis process. It uses a multi-linear PCA face model (C. Cao, Weng, Zhou, et al., 2013; Paysan et al., 2009; X. Zhu et al., 2016) which has $n = 53$k vertices and *106*k faces:

$$M_{geo}(\alpha_{id}, \alpha_{exp}) = \bar{S} + S_{id} \cdot \alpha_{id} + S_{exp} \cdot \alpha_{exp} \tag{4.1}$$

$$M_{alb}(\alpha_{alb}) = \bar{T} + T \cdot \alpha_{alb} \tag{4.2}$$

This parametric face model has three dimensions, where identity and expression represent the facial geometry - $M_{geo}(\cdot) \in \mathbb{R}^{3n}$ and the third dimension represents the skin reflectance (albedo) - $M_{alb}(\cdot) \in \mathbb{R}^{3n}$. It assumes that the geometry and albedo obey a multivariate normal distribution centered at the average shape $\bar{S} = \bar{S}_{id} + \bar{S}_{exp} \in \mathbb{R}^{3n}$ ($\bar{S}_{id} \in \mathbb{R}^{3n}$ represents the average shape's identity part and $\bar{S}_{exp} \in \mathbb{R}^{3n}$ represents the expression part) and reflectance $\bar{T} \in \mathbb{R}^{3n}$. The corresponding bases are $S_{id} \in \mathbb{R}^{3n \times 99}$, $S_{exp} \in \mathbb{R}^{3n \times 29}$ and $T \in \mathbb{R}^{3n \times 99}$. Standard deviations are $\sigma_{id} \in \mathbb{R}^{99}$, $\sigma_{exp} \in \mathbb{R}^{29}$ and $\sigma_{alb} \in \mathbb{R}^{99}$. These bases and standard deviations are obtained by applying Principal Component Analysis on a set of 3D face scans (Paysan et al., 2009) and assumed to be known in 3D face reconstruction.

The unknown variables in the aforementioned parametric face model are $\alpha_{id} \in \mathbb{R}^{99}$, $\alpha_{exp} \in \mathbb{R}^{29}$ and $\alpha_{alb} \in \mathbb{R}^{99}$. They control the target face's 3D geometry and skin reflectance.

The face is assumed to have the Lambertian surface reflectance and the illumination is modelled with a second order Spherical Harmonics (SH) denoted by $L \in \mathbb{R}^{27}$ (Basri & Jacobs, 2003). A face image I_{syn} is synthesized by rasterizing the parametric face model under a rigid transformation (R, t) and a perspective projection $\Pi_P(M_{geo})$ with the camera parameters K.

In an analysis-through-synthesis loop, face model and rendering parameters are optimized mainly along the direction of generating a face image as close as the input image. The objective function is formulated as:

$$E(\mathcal{P}) = w_{col}E_{col}(\mathcal{P}) + w_{lan}E_{lan}(\mathcal{P}) + w_{reg}E_{reg}(\mathcal{P}) \tag{4.3}$$

where $\mathcal{P} = \{\alpha_{id}, \alpha_{exp}, \alpha_{alb}, R, t, K, L\}$ is the collection of facial geometry and reflectance parameters $(\alpha_{id}, \alpha_{exp}, \alpha_{alb})$, rigid motion parameters (R, t), camera parameters K and illumination parameters L. $w_{col} = 1$, $w_{lan} = 10$ and $w_{reg} = 2.5 \times 10^{-5}$ are empirical weights to balance three energy terms - $E_{col}(\mathcal{P})$, $E_{lan}(\mathcal{P})$ and $E_{reg}(\mathcal{P})$.

The photo-consistency term E_{col} measures the colour distance between the synthesized face image and the input image:

$$E_{col}(\mathcal{P}) = \frac{1}{|V_I|}\sum_{v_I \in V_I}\left\|I_{syn}(v_I) - I_{in}(v_I)\right\|_2 \tag{4.4}$$

$$I_{syn}(v_I) = \left[I_r^v, I_g^v, I_b^v\right]^T$$

$$I_{ch}^v = M_{alb,\,ch}^v \cdot \sum_{i=1}^{9}\gamma_{i,ch}y_i\left(\boldsymbol{n}(v)\right),\ ch \in \{r, g, b\}$$

$$v_I = \Pi_P(Rv + t)$$

where I_{in} is the input image and $v_I \in V_I$ denote all visible pixel locations in the synthesized image I_{syn}, v_I is obtained by projecting the visible 3D vertex v on the face mesh onto the image plane. Its pixel value $I_{syn}(v_I)$ is assigned with v's

texture value - I_{ch}^{v}, wherey_i is the ith SH basis function, $\gamma_{i,ch}$ is the corresponding SH coefficient of a specific color channel – ch, $\boldsymbol{n}(v)$ is the normal of v, and $M_{alb,\,ch}^{v}$ is the albedo value of v in channel ch. The landmark-fitting term E_{lan} enforces a constraint to the reconstructed facial geometry according to some fiducial facial points, namely projected 3D landmarks, which should align to the corresponding landmarks on the input face image as accurate as possible:

$$E_{lan}(\mathcal{P}) = \frac{1}{|F|}\sum_{f_i \in F}\|f_i - \Pi_P(Rv_i + t)\|_2^2 \tag{4.5}$$

f_i is a 2D facial landmark detected on the input face image from my implementation of (Xiong & De la Torre, 2013). To ensure the plausibility of the reconstructed 3D face, a statistical regularization term is used to restrict face model parameters to a reasonable range:

$$E_{reg}(\mathcal{P}) = \sum_{i=1}^{99}\left[\left(\frac{\alpha_{id,i}}{\sigma_{id,i}}\right)^2 + \left(\frac{\alpha_{alb,i}}{\sigma_{alb,i}}\right)^2\right] + \sum_{i=1}^{29}\left(\frac{\alpha_{exp,i}}{\sigma_{exp,i}}\right)^2 \tag{4.6}$$

The objective function is transformed with the method of Iteratively Reweighted Least Squares (IRLS) and optimized using a Gauss-Newton (GN) solver. In my implementation, the optimization converges within 7, 5 and 3 GN steps from the coarsest level to the finest level of a three-level image pyramid (see Fig. 4.5). For generating personalized blendshapes, the expression component will be removed from the reconstructed 3D face. Please also note that the reconstructed facial texture presented in this work only keeps the estimated albedo and is rendered with a default lighting (see Fig. 4.6). The estimated lighting is discarded because it only models the lighting of the input face image, while the reconstructed 3D face should be rendered with the lighting of the virtual space for a more realistic face embodiment.

4.4.2 Personalized Blendshapes Generation

To get a digital face embodiment with AUs, the system adopts a robust non-rigid registration algorithm (S. Zhang et al., 2017) and a series of template blendshapes

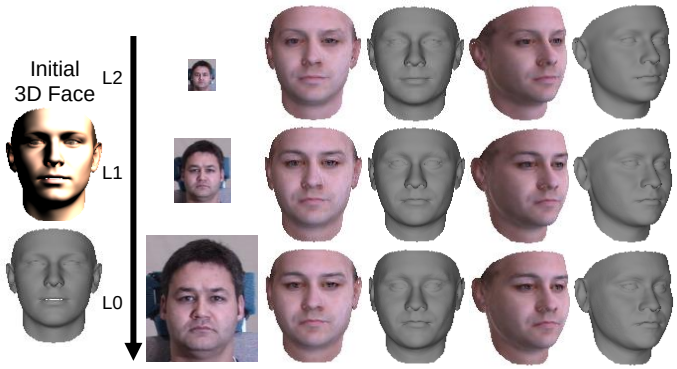


Figure 4.5: Reconstruction results from each level of the image pyramid. The initial 3D face is set with mean identity, neutral pose, mean albedo and rendered with the lighting model whose 27 Spherical Harmonics (SH) coefficients are all set with 1. The facial texture here is rendered with estimated albedo and lighting. Face images have been cropped and resized for better presentation. The original face image is 640*480.

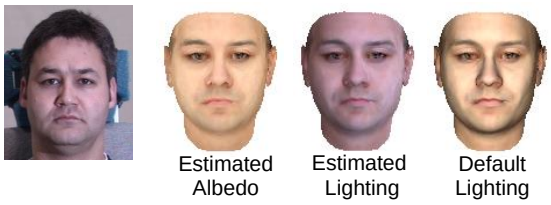


Figure 4.6: Reconstructed facial texture. The right two images show faces rendered with the estimated lighting and the default lighting used in this work.

(H. Yu et al., 2012). The template blendshapes are based on 3D scans of three female FACS certified actors using a 4D stereo imaging system (Imaging, 2020). Each actor performed 20 to 30 AUs, providing a total of 37 AUs (counting lateralizations) as well as a neutral face. The template mesh consists of 4,735 vertices and 8,760 faces. In this work, I use seven AUs (AU1, AU2, AU4, AU6, AU12, AU25 and AU43) for association with and prediction from the available EMG signals. I calculate the mean of all actor's meshes to get a more general template (see Fig. 4.4).

The non-rigid registration conforms the template neutral face (source point cloud) to the reconstructed 3D face (target point cloud). A coupled global and local deformation is applied to the vertex v_i on the source point cloud:

$$\tilde{\mathrm{v}}_i = \Phi_{local} \circ \Phi_{global}(\mathrm{v}_i) \tag{4.7}$$

A rotation matrix R relative to the centre-of-mass m and a translation vector t define the global rigid deformation:

$$\Phi_{global}(\mathrm{v}_i) = R(\mathrm{v}_i - m) + m + t \tag{4.8}$$

The non-rigid deformation is defined by a set of deviation vector d_i:

$$\Phi_{local}(\mathrm{v}_i) = \mathrm{v}_i + \mathrm{d}_i \tag{4.9}$$

For each vertex $\mathrm{v}_i$on the source point cloud, I associate a corresponding position $\mathrm{c}(\mathrm{v}_i)$ on the target which is initialized with a closest point computation and updated iteratively within the optimization. The non-rigid registration can hence be casted as an unconstrained energy minimization problem with unknowns $\mathcal{K} = \{R, t, \mathbf{D}, \mathrm{c}\}$, where $\mathbf{D} = \{\mathrm{d}_i\}$. The objective function is formulated as:

$$E(\mathcal{K}) = w_{fit}E_{fit}(\mathcal{K}) + w_{smooth}E_{smooth}(\mathcal{K}) \tag{4.10}$$

The weights w_{fit} and w_{smooth} compensate for different scales of the energy terms.

E_{fit} measures the corresponding point distance between the source point cloud and the target point cloud:

$$E_{fit}(\mathcal{K}) = \sum_{i=1}^{n} w_{conf,i}^2 \|\tilde{\mathrm{v}}_i - \mathrm{c}(\mathrm{v}_i)\|_2^2 + \sum_{i=1}^{n} \left(1 - w_{conf,i}^2\right)^2 \tag{4.11}$$

where n is the number of correspondences and $w_{conf,i}$ is the confidence of the each correspondence. And $w_{conf,i}$ close to one indicates a reliable correspondence, while $w_{conf,i}$ close to zero indicates that no proper correspondence is found. Source vertices without valid correspondence are excluded from the optimization process. The texture of the reconstructed face is transferred to the neutral template using the correspondence as well. To enhance the surface smoothness, an energy term enforcing small changes of point neighbourhoods and triangle areas is augmented to the objective function:

$$E_{smooth}(\mathcal{K}) = E_{neigh}(\mathcal{K}) + E_{area}(\mathcal{K}) \tag{4.12}$$

$$E_{neigh}(\mathcal{K}) = \sum_{i=1}^{n} \sum_{j=\mathcal{N}(i)} \left(\left\| \tilde{\mathrm{v}}_i - \tilde{\mathrm{v}}_j \right\|_2 - \left\| \mathrm{v}_i - \mathrm{v}_j \right\|_2 \right)^2 \quad (4.13)$$

$$E_{area}(\mathcal{K}) = \sum_{i=1}^{n} \left(\mathcal{A}(\tilde{\mathrm{v}}_i) - \mathcal{A}(\mathrm{v}_i) \right)^2 \quad (4.14)$$

where $\mathcal{N}(i)$ is the one-ring neighbourhood and $\mathcal{A}(\mathrm{v}_i)$ is the summing area of triangles attached to v_i on the source mesh.

I use the Levenberg-Marquardt algorithm to solve the non-linear least squares problem above. After obtaining the deformed neutral template, I calculate deviations between the original neutral template and template blendshapes. These deviations are then transferred to the deformed neutral template to generate personalized blendshapes. The texture of the deformed neutral template is transferred at the same time, providing fully-textured personalized blendshapes.

4.5 Facial Expression from EMG Signals

Facial expressions are results of facial muscle movements. The pioneering hardware solution – Faceteq offers an efficient way to sense facial muscle contractions through eight integrated EMG sensors placed on the ESFP. However, mapping EMG signals to facial expressions is nontrivial. To this end, 15 subjects are recruited and the Faceteq interface is used for data collection and analysis. All facial expressions are defined according to action units (AUs), which enables a convincing facial expression recovery in subsequent steps.

4.5.1 Data Collection

In this study, 15 volunteers are recruited (11 male and 4 female), aged from 21 to 52 years old (Mean: 31.93, Std: 12.75). Each participant is asked to perform five common facial expressions (see Fig. 4.7) – closed mouth smile, eye closure, forehead wrinkle, frown and open mouth smile, while wearing the prototype of the EMG-based facial sensing interface. All the facial expressions are defined with AU combinations following the work (Ekman & Rosenberg, 1997): AU6, 12 for closed mouth smile, AU43 for eye closure, AU1, 2 for forehead wrinkle, AU4 for

Figure 4.7: AU-coded facial expressions studied in this work. From left to right: forehead wrinkle, frown, eye closure, close mouth smile, and open mouth smile.

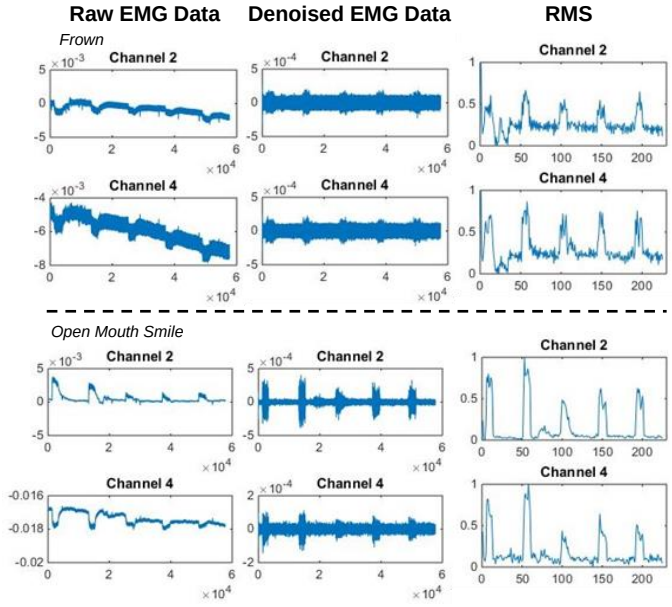


Figure 4.8: EMG signals and RMS features. The left two columns compare raw and denoised EMG data. The right column shows the RMS feature extracted from the denoised EMG data within a 256-msec time window.

frown and the combination of AU6, 12 and 25 for open mouth smile. Eight surface dry EMG electrodes are placed symmetrically on the left and right side of the Faceteq interface, providing eight-channel EMG signals. These EMG electrodes monitor activations of specific facial muscles (see Fig. 4.3), including Zygomaticus major (channel 1&2), Frontalis (channel 3&4), Orbicularis oculi (channel 5&6) and Corrugator (channel 7&8).

An audio track is used to instruct the subject to make facial expressions or return to the neutral pose. Each facial expression is repeated ten times with each lasting for two seconds. There is a ten-second rest between two adjacent repetitions.

EMG signals are recorded at 1kHz sampling rate, resulting in $2 \times 10 \times 1{,}000 = 20{,}000$ EMG samples for each facial expression of each subject in theory. The actual EMG sample amount may drift around the estimated value as there is latency in starting or ending the facial expression for each subject.

4.5.2 EMG Signal Pre-Processing

Facial EMG signals have small amplitude and can be easily interfered by various external or internal factors, such as motion artefacts, incorrect sensor placement and environmental noise. I therefore use multiple filters to clean the raw EMG data. A baseline correction on raw EMG signals is first adopted to remove mean values and the linear trend. To eliminate artefacts such as the line interference introduced by electrical devices, the Notch filter is then applied to remove the 50Hz component and its harmonics up to 350Hz. Signals are further passed through a band-pass filter retaining components from 30 to 450Hz. Finally, the eight-channel clean EMG signals can be obtained (see Fig. 4.8).

4.5.3 Feature Extraction

To reduce the dimensionality of data and extract the most informative segments, it is crucial to compress EMG signals along the time axis. Generally, EMG signals are partitioned into temporal segments of the same length, from where features are extracted. Long segments can suppress bias and variance of the feature, however, they may fail to reach the efficiency requirement (Oskoei & Hu, 2007). Some recent works report that using segments with 256 msecs length is a good trade-off between the feature effectiveness and the overall processing efficiency (Hamedi et al., 2016; Rezazadeh et al., 2011). I follow the setting of (Hamedi et al., 2016) by segmenting the pre-processed EMG signals into non-overlapping 256-msec pieces.

Root Mean Square (RMS) is one of the representative time-domain features and has been widely used for analysing the contraction of facial muscle. With the hypothesis of the Gaussian random process, RMS provides the maximum likelihood estimation of EMG amplitude when a facial muscle is under constant

force and non-fatiguing contraction. According to a recent survey (Hamedi et al., 2016), RMS shows superiority against the other time-domain features. I hence extract RMS from each 256-msec segment of EMG signals:

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{4.15}$$

where $N = 256$, x_n is an EMG sample within the 256-msec segment.

4.5.4 Facial Expression Prediction

Multiple external or internal factors such as EMG electrode drift, individual variances or muscle fatigue, usually result in a large variation of the EMG pattern even for the same facial expression. Hence, a robust learning algorithm is required for mapping EMG features to facial expressions accurately. The influence of the classifier to the final prediction performance has been much studied. A recent study (Hamedi et al., 2016) compares 14 classifiers and reports that Least-square Support Vector Machine (LS-SVM), Regularized Discriminative Analysis (RDA), Normal Density Discriminant Function (NDDF) and Maximum-likelihood (ML) estimation provide a much higher classification accuracy against the other classifiers. ANOVA analysis shows that there is no significant difference among the classification performance of the top four classifiers (Hamedi et al., 2015). In this work, I choose LS-SVM as the classifier and adopt the libSVM (Chang & Lin, 2011) framework to train the multiclass LS-SVM.

4.6 Basic Emotions Prediction from AUs

Existing rule-based methods (Ozuysal et al., 2007; Valstar & Pantic, 2006) cannot be extended to scenarios where observed AUs are not included in the established heuristic rules. Restricted by the number of EMG sensors applied and the range of facial expressions covered by the collected database, the developed system outputs specified AUs. It makes previous rule-based methods infeasible. To address this problem, I propose to use the fern classifier to model the relationship between AUs

and six basic emotions. Specifically, my target is to learn the posterior probability of emotions given occurrences of AUs. It is a typical Bayesian classification problem that can be solved efficiently with the fern classifier.

Fern classifier has been successfully applied in image key-points recognition (Ozuysal et al., 2007). Each fern is a composition of a small set of features and a series of binary tests on these features. It returns the probability that a sample belongs to a class. Outputs from all ferns are then combined together in a Naive Bayesian way. In my task, six basic emotions - anger, disgust, fear, happiness, sadness and surprise are treated as classes, while the occurrence of AU is a binary feature. Since the feature pool consists of limited AUs, there is no need to partition it into groups of features. One fern is sufficient to learn the class-conditional distribution in my case. Let $c_i, i = 1, \dots, H$ be the set of class (emotion) and $f_j, j = 1, \dots, N$ be the set of binary feature (AU occurrence), we are looking for:

$$\hat{c}_i = \underset{c_i}{\operatorname{argmax}}\, P(C = c_i \mid f_1, f_2, \dots, f_N) \tag{4.16}$$

$$f_j = \begin{cases} 1 & if\ AU_j\ occurs \\ 0 & otherwise \end{cases}$$

where C represents the class. With Bayes' theorem, we can have:

$$P(C = c_i \mid f_1, f_2, \dots, f_N) = \frac{P(f_1, f_2, \dots, f_N \mid C = c_i) P(C = c_i)}{P(f_1, f_2, \dots, f_N)} \tag{4.17}$$

$$P(f_1, f_2, \dots, f_N) = \sum_{i=1}^{H} P(f_1, f_2, \dots, f_N \mid C = c_i) P(C = c_i)$$

where $P(C = c_i)$ is the prior probability of emotion.

I build the fern classifier on two benchmark facial expression databases - CK+ (Lucey et al., 2010) and EmotioNet (Fabian Benitez-Quiroz et al., 2016) that contain both AU and basic emotion labels.

CK+ involves 123 subjects who are instructed to perform 23 facial expressions forming a database of 593 image sequences. Each sequence incorporates the onset (the neutral face) to peak formation of the facial expression. The peak frame of the facial expression is coded with AU and seven basic emotion labels (anger, contempt, disgust, fear, happiness, sadness and surprise) which are further rectified

according to the FACS manual (Ekman & Rosenberg, 1997). Overall, CK+ offers 327 samples with both AU and basic emotion labels for this study. As contempt is beyond the scope of this study, I remove samples labelled as contempt, leaving 309 samples in total for subsequent analysis.

EmotioNet consists of a million in-the-wild images of facial expressions in which 975,000 images are made available to the public. Within the released database, there are 950,000 images automatically annotated with AUs and AU intensities. The remaining 25,000 are manually annotated with AUs by qualified coders. A small part of these images (2,479 images) are labelled with one of 16 compound emotions defined in (Du et al., 2014) based on AU combinations. Since this study only considers six basic emotions, I relabelled images according to their compound emotion labels. For example, if an image has been annotated as happily surprised, I categorize it into both happiness and surprise. In total, there are 3,581 samples available with AU and basic emotion labels.

CK+ and EmotioNet, covering a wide range of AU-emotion relations, are appropriate for statistical analysis. Experimental results demonstrate the proposed method is able to give valuable emotion information when only limited AUs are available. This function is hence incorporated into the proposed system to assist the VR HMD user to understand the other users' emotions in the virtual world.

4.7 Results and Analysis

The proposed system is developed to enable realistic facial expression reconstruction for the VR user wearing a HMD. It has been validated on 13 subjects. Each of the three principal system parts – face embodiment construction, facial expression prediction and basic emotions estimation, has been carefully evaluated. In the following, I will report experimental results of each part and the overall system performance afterwards.

4.7.1 Face Embodiment Construction

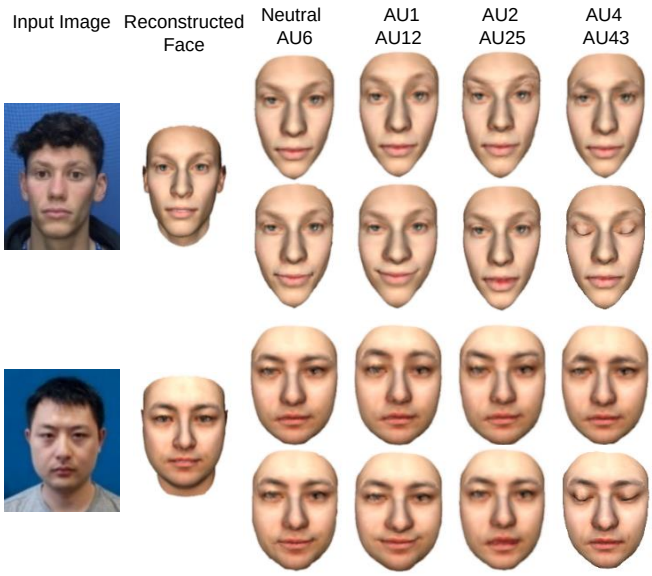


Figure 4.9: Realistic face embodiment generation from a single image. The reconstructed face shown above has discarded facial expression components.

I evaluate the robustness of this part using face images from various subjects. The results are shown in Fig. 4.9. With a single face image of the user, the developed system is able to reconstruct a fully textured 3D face and generate vivid AU-based blendshapes.

Since each AU originates from 3D scans of FACS certified actors, the generated blendshapes form a solid basis for natural facial expression composition. To enable deformation transfer between AU and the neutral face, I remove facial expression components from the reconstructed 3D face.

4.7.2 Facial Expression Prediction

Predicting facial expressions from dry EMG signals is not easy as the raw signals are quite noisy. I thus apply the aforementioned multiple processing steps to clean the raw signals. Fig. 4.8 compares the raw EMG signal and the denoised signal. EMG signals from channel 2 and channel 4 when collecting data for frown and open mouth smile are plotted for illustration. After obtaining clean EMG signals, I extract the RMS feature within a 256-msec time segment using Eq. 4.15 (see Fig.

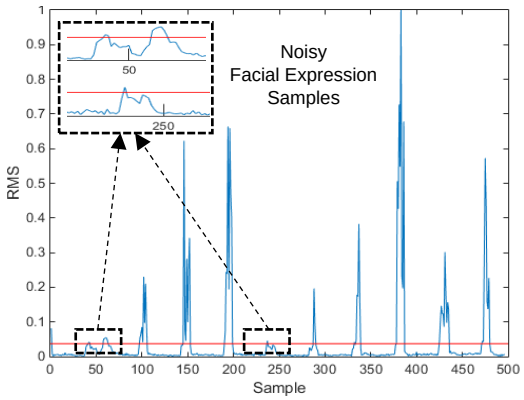


Figure 4.10: Representative RMS values of channel-1 EMG signals of the closed mouth smile. The red line indicates the reference RMS level which is used to differentiate facial expression samples from neutral samples. The black dashed box indicates noisy facial expression samples that were labeled as neutral.

4.8 for example). As the instructional audio track was started manually, there are some variability in the time recorded and therefore the data. It also had a very short time delay to make or stop the facial expression once the subject heard the prompt tone. Therefore, it is infeasible to accurately annotate the RMS samples with correct facial expression labels according to the timestamp. To separate facial expression samples from neutral samples, I calculate a RMS threshold - $mean(RMS) + n \times std(RMS)$ (shown as the red line in Fig. 4.10) for each EMG channel (Hamedi et al., 2011), where n is an empirical value set manually according to each RMS curve. In my experiment, $n = 0.25$ works well for all tests. Samples whose RMS values are above at least one threshold line are annotated with the specific facial expression label, while the others are annotated as neutral. This process extracts the most significant facial expression samples while filtering out samples with insignificant RMS features which were probably caused by the distraction of the subject or other noise-related interferences.

The overall distribution of facial expression RMS samples is listed in Table 4.1. To balance the neutral and facial expression samples in the data set, I randomly select the same number of neutral samples as the facial expression samples for each facial expression of each subject. I conduct two experiments to validate the

Table 4.1: The distribution of labelled facial expression RMS samples.

	CMS	EC	FW	FR	OMS
Sample Number	2061	2416	1940	2483	1853

CMS: closed mouth smile EC: eye closure FR: frown
FW: forehead wrinkle OMS: open mouth smile

accuracy of facial expression prediction from RMS features. For both experiments, I perform max-min normalization for all the RMS features of the data using the max and min RMS values of the training data (Halaki & Ginn, 2012). The max and min RMS values are chosen for each EMG channel and each facial expression.

First, I use the whole dataset and randomly select 80% data for training and the left 20% data for testing (Hamedi et al., 2015, 2016). The LS-SVM with the RBF kernel is adopted for the target multi-class classifier. 10-fold cross validation is applied to select the hyper-parameters C and $gamma$ which are set as 2 and 32 respectively in my experiment. The classification accuracy on the whole testing set is 86.77%. I also calculate the classification accuracy for each subject or each facial expression in the testing set (see E1 in Table 4.2). The average classification accuracy across subjects is 86.27% with a standard deviation of 0.0644. The average classification accuracy across expressions is 84.74% with a standard deviation of 0.1867.

To further validate the generality of the facial expression classifier to EMG signals from new subjects, I apply the leave-one-out validation. Specifically, I leave one subject's data for testing while using the data from the other subjects for training. I repeat the same process for each subject, which results in 15 different classifiers. The performance of the classifier is shown in E2 in Table 4.2. The mean classification accuracy for 15 classifiers is 75.8% with a standard deviation of 0.1033. To validate the effectiveness of the EMG signal denoising step and the RMS feature, I conduct the same experiment as in E2 on denoised EMG signals and RMS features extracted from raw EMG signals. When the denoised EMG signals are fed directly into the classifier, the classification accuracy declines

Table 4.2: Facial expression recognition accuracy from RMS features.

Classification Accuracy							
	E2						E1
	CMS	EC	FW	FR	OMS	ALL	ALL
S1	0.16	0.24	0.96	1.00	0.01	0.60	0.73
S2	0.89	0.45	0.97	0.99	0.43	0.77	0.89
S3	0.28	0.77	0.98	1.00	0.66	0.74	0.88
S4	0.86	0.43	0.96	1.00	0.16	0.73	0.84
S5	0.42	0.62	0.95	0.82	0.86	0.68	0.94
S6	0.75	0.85	0.97	1.00	0.16	0.82	0.94
S7	0.81	0.84	1.00	1.00	0.24	0.76	0.78
S8	0.55	0.93	0.99	1.00	0.74	0.87	0.90
S9	0.38	0.96	0.99	0.80	0.65	0.82	0.89
S10	0.30	0.97	0.99	1.00	0.92	0.87	0.91
S11	0.97	0.88	1.00	1.00	0.28	0.80	0.91
S12	0.87	0.32	0.98	1.00	0.18	0.70	0.78
S13	0.65	0.47	0.73	0.59	0.54	0.51	0.79
S14	0.71	0.91	0.99	0.98	0.71	0.87	0.90
S15	0.95	0.87	0.97	1.00	0.23	0.83	0.86
E1	0.79	0.91	0.99	0.99	0.55		

Note: 1) E1-E2 represent experiments; 2) The results of E2 are in orange and the results of E1 are in purple. 3) S1-S15 represent subjects; 4) ALL is the classification accuracy on all the testing data, including neutral samples.

significantly, with a mean of 20.74% and a standard deviation of 0.0316. The results also show that facial expressions cannot be accurately predicted from RMS features extracted from raw EMG signals (mean: 25.16%, standard deviation: 0.2016).

As shown in Table 4.2, both experiments show high classification accuracy for FW and FR, while lower accuracy for CMS and OMS. It can also be found that the classification accuracy varies from one subject to another. For example, S8 and S10 achieve high accuracy in both experiments, while S1 and S13 have a much lower recognition rate, even when their data was included in training in E1. This is probably due to EMG signals from some subjects are noisier or contain patterns that are quite different from EMG signals of the other subjects. This can be remedied by collecting data from more subjects and improving the data capture process.

4.7.3 Basic Emotion Estimation

All facial expressions involved in this study have been coded in AUs, providing a detailed and anatomic description of facial expressions. AUs describing the physical appearance of facial display offer valuable clues for predicting six basic emotions. The developed system outputs AUs – AU1, AU2, AU4, AU6, AU12, AU25 and AU43 focusing on the ESFPs, which makes previous rule-based methods intractable. To get an insight into the HMD user's internal emotional states, I build the relationship between AUs and six basic emotions with a probabilistic model – the fern classifier. Following Bayes' theorem, the model estimates the posterior probability of a basic emotion when observing a specific group of AUs.

The probabilistic model is learned from two benchmark FACS-annotated facial expression databases – CK+ (Lucey et al., 2010) and EmotioNet (Fabian Benitez-Quiroz et al., 2016). There are only two samples observing AU43 in CK+, while AU43 is not used when defining compound emotion category in EmotioNet. I hence discard AU43 when estimating basic emotions.

As shown in Table 4.3, the number of samples belonging to each emotional category varies from each other. If the prior probability of an emotion is regarded as the proportion of samples belonging to it, the posterior probability of emotion given an AU combination will become the proportion of samples coded in the current AU combination within the whole database (see Eq. 4.17). This will cause large deviations when calculating probabilities. I therefore assumed that basic emotions have identical prior probabilities.

After applying Eq. 4.17, the probabilities of basic emotions given the occurrence of AU for both databases are obtained (see Table 4.3). Table 4.3 includes observed combinations of AUs that are used to define facial expressions in this work. From the results, I found the following phenomena:

1) Emotions can be expressed in various forms of facial expressions. As can be seen from the table, when none of AUs specified in this work occur, emotions

Table 4.3: The probability of emotion given AUs learned from CK+ and EmotioNet.

CK+ (309)						
AU code AU25_AU12_AU6 _AU4_AU2_AU1	Anger (45)	Disgust (59)	Fear (25)	Happiness (69)	Sadness (28)	Surprise (83)
'000000'	0.0655	0.8990	0	0	0	0.0355
'000001'	0	0	0	0	1	0
'000011' - AU1, 2 (3)	0	0	0	0	1 (3)	0
'000100' - AU4 (54)	0.6839 (35)	0.2533 (17)	0	0	0.0628 (2)	0
'000101'	0	0	0.0618	0	0.9382	0
'000111'	0	0	0.2188	0	0.7183	0
'001000'	0.7387	0.1409	0	0.1204	0	0
'001100'	0.2875	0.7125	0	0	0	0
'010100'	1	0	0	0	0	0
'011000' - AU6, 12 (2)	0	0.5391 (1)	0	0.4609 (1)	0	0
'100000'	0	0.3897	0	0.3332	0	0.2770
'100011'	0	0	0.1471	0	0	0.8529
'100100'	0	0.3610	0.6390	0	0	0
'100101'	0	0	1	0	0	0
'100111'	0	0	0.9300	0	0	0.0700
'110000'	0	0.3690	0	0.6310	0	0
'111000' - AU6, 12, 25 (64)	0	0	0	1 (64)	0	0
'111101'	0	0	1	0	0	0
EmotioNet (3,581)						
	Anger (289)	Disgust (977)	Fear (150)	Happiness (1536)	Sadness (359)	Surprise (270)
'000000'	0	1	0	0	0	0
'000100' - AU4 (760)	0.3268 (173)	0.1274 (228)	0	0	0.5459 (359)	0
'100011'	0	0	0.4348	0	0	0.5652
'100100'	0.4099	0	0.3404	0	0	0.2496
'100101'	0	0	1	0	0	0
'110000'	0	0.4274	0	0.5726	0	0
'110011'	0	0	0	0.1495	0	0.8505

Note: 1) AU code: '1' indicates the AU occurs, '0' indicates the AU doesn't occur; 2) the number in parentheses denotes the amount of samples belonging to the category

such as disgust, anger, surprise can still be observed. On the other side, a combination of AUs can describe several basic emotions. For instance, in CK+, AU6 and AU12 indicate both disgust and happiness, while the results of EmotioNet show that AU4 and AU25 probably can be observed for anger, fear and surprise. Most combinations of AUs used in this work, namely AU6,12, AU1,2 and AU4 etc. are not discriminative for predicting an emotion category.

2) A few AUs or combinations of AUs show stronger links to emotions than the others. In CK+, AU1,4 indicates a high probability (0.9382) of sadness, while AU6,12,25 indicates a high probability (1.0) of happiness. In both CK+ and EmotioNet, AU4 normally indicates a negative emotion, such as anger, disgust and sadness. It can be found that the specific probability of each emotion when AU4 occurs differs between CK+ and EmotioNet. It is mainly due to the variety of facial expressions of emotions, which leads to a single database only covering a limited range of AU-emotion relationships. This also explains the first phenomena. Some other AUs or AU combinations, e.g. AU25 can be found in all six basic emotions.

To further verify the second phenomena, I calculate the discriminative power of each AU to an emotion (Velusamy et al., 2011):

$$D = P(A_j|E_i) - P(A_j|\bar{E}_i) \tag{4.18}$$

where $P(A_j|E_i)$ is the probability of observing AU A_j when emotion E_i occurs, and $P(A_j|\bar{E}_i)$ is the probability of observing A_j when E_i doesn't occur. D measures the relationship between the AU and the emotion. D close to -1 represents a strong negative correlation, while D close to 1 represents a strong positive correlation. I generate a correlation matrix of AUs and emotions expressed with the discriminative power D from CK+. Each D is normalized across all the AUs for each of the emotions. The original relation matrix contains 35 AUs and 7 emotions. I remove most matrix components while only keeping AUs and six basic emotions studied in this work. As shown in Fig. 4.11, AU1 associates closer to fear, sadness and surprise, while AU6 and AU12 have a distinctive connection with happiness. AU4 shows a closer relation with anger, fear and sadness. AU2 links closely to surprise. Consequently, the correlation matrix demonstrates the emotional saliency of the studied AUs and is consistent with the previous probabilistic model.

Overall, the above probabilistic analysis takes a deep insight into the relationship between AUs and six basic emotions. The built probabilistic model

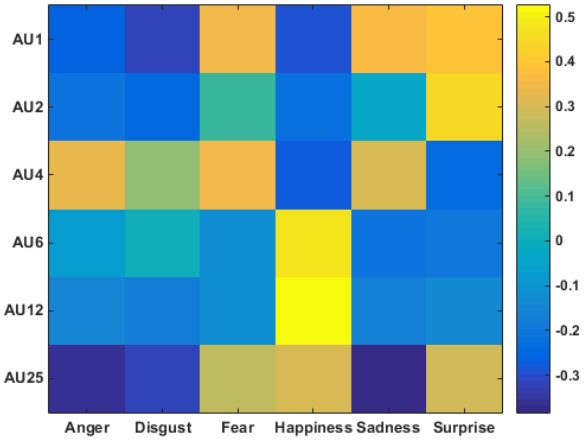


Figure 4.11: Correlations between AUs and basic emotions. The value close to 1 represents a strong positive correlation, while the value close to -1 represents a strong negative correlation.

can provide useful emotional information when only limited AUs are available.

4.7.4 Full System Evaluation

I first test the system with the prototype of Faceteq which was used to collect the EMG data of specified facial expressions. Since the prototype can be used independently from a VR HMD and hence doesn't occlude the user's principal face region, it provides direct comparisons between the reconstructed 3D facial expression and the ground truth. Example results are shown in Fig. 4.12. The facial sensing hardware can detect the user's facial expression through eight integrated EMG sensors. The facial expression is then mapped onto a realistic face embodiment of the user.

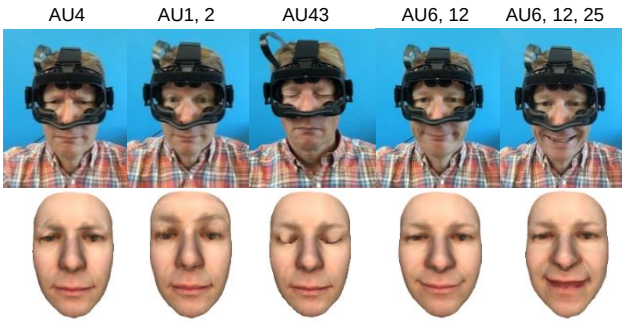


Figure 4.12: Facial expressions sensed and reconstructed with the proposed system when the user was wearing the Faceteq prototype.

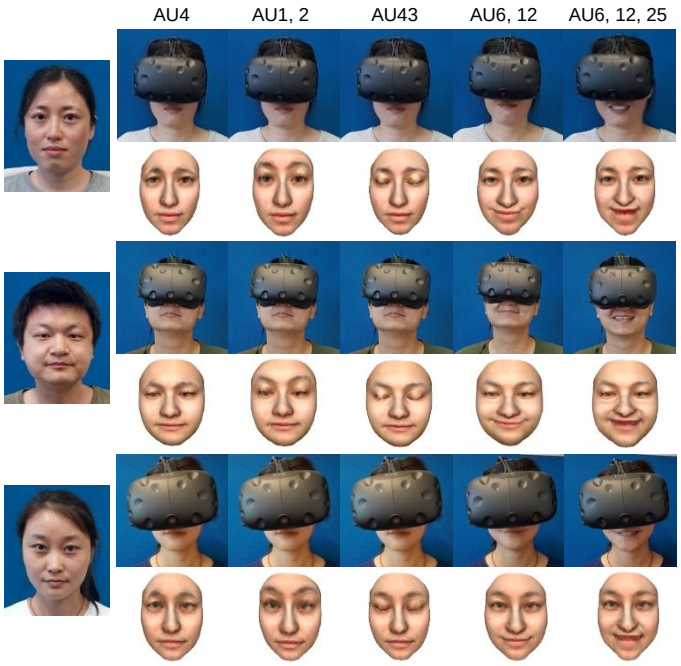


Figure 4.13: Facial expressions sensed and reconstructed with the proposed system when the user was wearing the VR HMD integrated with the Faceteq.

Then, I do validation tests with the Faceteq and the VR HMD. As the EMG-based facial sensing interface has been designed to softly enfold the wearer's face, the type of HMD will not affect the scale of the EMG data. During the test, each subject was asked to perform facial expressions specified in this study when wearing a commercial VR HMD attached with the Faceteq hardware. Fig. 4.13 shows example results. With the learned fern classifier, a probabilistic model between AUs and six basic emotions can be further obtained, which is important to VR applications.

As shown in Fig. 4.14, existing systems apply different hardware and avatars to capture and exhibit the user's facial performance. This makes it difficult to compare systems on quantitative metrics such as the 3D facial expression reconstruction accuracy. Moreover, those systems are developed for capturing different facial expressions and emotions, which also prohibits the comparison on the expression/emotion recognition rate. As a result, visual comparison is the only feasible way at this stage. Whereas it has limitations, it can intuitively reflect the

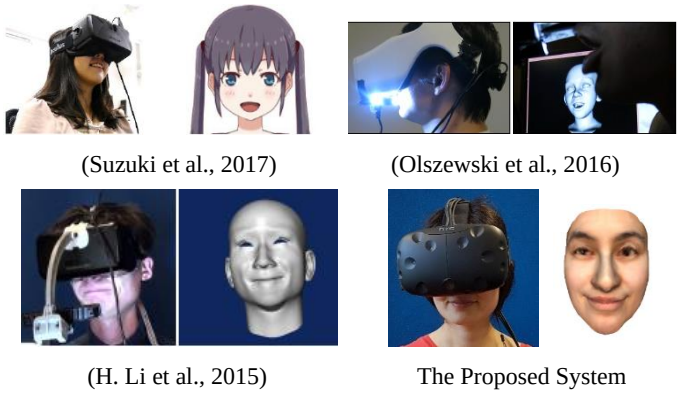


Figure 4.14: Comparison with other similar systems from (H. Li et al., 2015; Olszewski et al., 2016; Suzuki et al., 2017). Avatars in (H. Li et al., 2015) and (Olszewski et al., 2016) capture the user's facial movements with a clumsy RGB/RGBD camera attached on the VR HMD, while not preserving the user's facial identity and texture. (Suzuki et al., 2017) simply uses a 2D cartoon image to represent the user's face.

visual quality of the captured facial expression. Fig. 4.14 visually compares the proposed system with three other representative systems in this field. It can be found that the proposed system can reconstruct a more realistic 3D face embodiment for the VR HMD user and doesn't require additional cameras to capture the user's facial movements.

4.7.5 Limitations

Since the biometric data collection step is labor-intensive and expensive, this work focuses on specific AU combinations on the ESFP. Future work could be extended to the detection of various AUs independently from EMG signals from a larger biometric database involving i) a wider selection of facial expression and ii) additional sensor inputs such as from an eye tracker. Then, any combination of these AUs obtained would be able to cover a wider range of facial expressions of emotion. Furthermore, the current system does not encode the intensity of facial expression. Both problems can be alleviated by collecting EMG data for single AUs with different scales of intensity. The current system mainly focuses on facial expressions displayed on the upper-half face, with the exception of AU12 and

AU25 which are sensed from the cheek. As such certain fundamental expressions occurring around the mouth area are ignored, in the future, I will also look to infer AUs associated with visemes during speech (Meng et al., 2017), such as AU24, AU25, AU26 and AU27.

Restricted by the required EMG signal processing, the system described above cannot attain a real-time performance. The original EMG data should be partitioned into a series of time sequences for feature extraction. Short sample windows lead to bias and variance in feature estimation, while long sample windows reduce the system efficiency. In the experiment, I applied a 256-msec time window, which means that only about four (3.9) frames can be output from the system in a second. This issue could potentially be alleviated by using shorter or overlapping time windows, or applying EMG sensors with a higher sampling rate.

The digital embodiment of the VR HMD user output from the proposed system is restricted to the frontal face region and still has significant room for improvement. A more compelling full head embodiment could be constructed by modelling hair (Hu et al., 2015), texture (Saito et al., 2017) and shape details (Huynh et al., 2018).

4.8 Conclusion

This chapter proposed a method and developed a prototype system that can sense and reconstruct the VR HMD user's 3D facial expression. The applied hardware component is portable and compatible with mainstream VR HMDs. It can detect facial muscle movements accurately with eight integrated EMG sensors placed on the ESFP. With a single face image, the system can reconstruct the user's fully textured 3D face and generate personalized AU-based blendshapes. Specifically, the system can capture AU-coded facial movements with integrated EMG sensors and a robust classifier learned from the data collected from 15 subjects. It can also provide useful emotional information for participants in the virtual world with a

novel probability model of AUs and six basic emotions built with the fern classifier. I believe the developed system can facilitate a wide range of VR applications, such as game, physical therapy and rehabilitation.

In future, I plan to equip the system with the ability to detect independent AUs and its intensity. A significant sized database consisting of AU and corresponding EMG signals will be created. I will extend to AUs associated with visemes during speech to cover more facial expressions. The system could potentially be improved to achieve real-time performance with additional biometric sensors or more efficient signal processing methods. Supplementary improvements could involve features such as hairs, texture and geometric details for a compelling full-head digital embodiment for VR applications.

Chapter 5

A Review on Automated Facial Nerve Function Assessment from Visual Face Capture

Foreword

After developing robust technical solutions for facial performance capture, the primary task becomes to apply and examine those solutions in real-world applications. Exploring new application avenues is thus essential and can in turn motivate the further development in capture methods. I noticed that existing face capture methods are mainly applied in content creation for entertainment, social media and human-machine interaction, the applications to facial biometrics for medical and health purpose only occupy a very small portion. Inspired by this finding, this chapter deeply investigates the largely unexplored problem – automated facial nerve function assessment from visual face capture for facial palsy management. It comprehensively reviews the principal studies in the field, and discusses the merits of existing assessment methods, the challenges and potential directions for further improvement. Most importantly, it identifies the significant role of monocular face capture approaches such as those developed in the previous three chapters in achieving purely automated, objective and accurate assessment. New directions of advancing the developed face capture approaches are also discussed, e.g. reconstructing and tracking asymmetric 2D/3D facial expressions for facial palsy patients.

The chapter is based on a published journal paper:

- **Lou, J.**, Yu, H., & Wang, F. Y. (2019). A review on automated facial nerve function assessment from visual face capture. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2), 488–497.

I conducted the investigation of the target problem and wrote up the review paper.

5.1 Introduction

Facial palsy is associated with a myriad of functional (A. Fattah et al., 2012) and psychosocial problems (Bradbury et al., 2006; Fu et al., 2011; VanSWEARINGEN et al., 1998; Walker et al., 2012) that erode foundations of the patient's health and daily life. It generally refers to the weakness of facial musculature innervated by the facial nerve. The main obstacle in facial palsy management is the lack of an effective tool to objectively assess and document facial nerve function, which is crucial to clearly understand the progression or resolution of disease, evaluate the outcomes of therapeutic interventions, and make an accurate prognosis and appropriate treatment plan.

A major part of facial nerve function refers to the motor function manifested by various facial muscle movements, and is visually observable with clear static or dynamic facial signs, e.g. resting symmetry, symmetry of voluntary movement and synkinesis. Facial nerve function assessment from these facial signs is hence an important means in clinical practice. With ubiquitous visual face capture – image/video, it is more widely accessible than those using obtrusive physical interventions such as electroneurography (ENoG) and electromyography (EMG). This motivated a branch of study (Gerós et al., 2016; Guarin et al., 2018; Z. Guo et al., 2017; Hadlock & Urban, 2012; Hsu et al., 2018; Ngo, Chen, et al., 2016; O'Reilly et al., 2010; Tzou et al., 2012; T. Wang et al., 2016) in this field to employ computational measures on biomedical visual face capture to objectively and quantitatively evaluate the facial nerve function. Such a solution is capable of automatically quantifying facial nerve function in millimetric precision (Guarin et al., 2018; Hadlock & Urban, 2012) or with semantic grades (Banks et al., 2015; Committee et al., 2009; House, 1983; Ross et al., 1996; Yanagihara, 1977) based on a machine learning model trained on clinician labelled data (Z. Guo et al., 2017; O'Reilly et al., 2010; T. Wang et al., 2016). This provides a highly efficient and cost-effective means whereby facial nerve function can be appraised in an objective manner. With the development of techniques in image processing,

computer vision and machine learning, especially those of 2D/3D face tracking (Gerós et al., 2016; Ngo, Chen, et al., 2016; Tzou et al., 2012) and feature learning (Hsu et al., 2018), the field of automated facial nerve function assessment has witnessed promising progress and developed various instruments in recent years. However, to date, none of these instruments has gained a wide clinical use. Their clinical effectiveness remains a big doubt mainly because of the limited data used for method development and validation. Meanwhile, important advancements in other areas, monocular 3D face tracking and face image synthesis for example, have not yet been fully utilized.

This study recaps the prerequisite knowledge of the facial nerve function and systematically reviews principal studies in automated facial nerve function assessment from visual face capture which contains rich physiological information (Samad et al., 2017), to provide readers with an overview of this critical research field and stimulate new ideas. It discusses existing challenges or problems and how they have been tackled so far, and indicates future directions. Particularly, it identifies the importance of monocular face capture approaches in solving the assessment problem, and discusses directions to further improve those approaches' performance. This introduces a promising avenue to apply and develop the face capture approaches proposed in the previous chapters. To the best of my knowledge, this is the first study of its kind to be reported so far, which is believed to benefit multiple groups of people, including researchers in visual face capture and clinical practitioners.

5.2 Review Methods

A systematic review of the English language literature published from 1977 to 2019 was performed from the resources of PubMed and Google Scholar according to agreed inclusion and exclusion criteria: **Inclusion** – 1) facial nerve function assessment from face images/videos using computational measures; 2) 2D/3D face analysis from visual face capture, including tracking, reconstruction, synthesis and

feature learning. **Exclusion** – 1) assessment from non-visual face capture, e.g. electroneurography and electromyography; 2) manual or subjective assessment methods; 3) non-English language.

Preliminary search was performed using key terms such as “facial”, “nerve”, “function”, “assessment”, “grading”, “palsy”, “paralysis”, “automated”, “automatic”, “computer”, “vision”, “machine”, “learning”, “image”, “video”, “processing”. These terms were manually grouped as key words for the search of titles which were then screened for potential relevance. 86 titles were searched and 47 articles were retrieved after carefully examining their abstracts and main contents with regard to relevance. Applying the same selection criteria, the searching scope was then extended to the bibliographies of all the selected publications for relevant reports that were not covered by database searching. This process yielded 15 more articles. The articles were grouped in terms of three aspects: 1) what kind of visual face capture was used such as still face image or dynamic facial expression image sequence, RGB or RGBD image; 2) whether provided quantification of static, dynamic and synkinetic facial features; 3) whether predicted semantic grades using machine learning techniques. These three aspects also grounded the subsequent categorization of different automated assessment methods.

In addition, 15 more articles that introduce facial nerve function, facial palsy, their clinical assessment, and clinical facial nerve grading scales were reviewed and summarized to briefly introduce the medical background of this review.

5.3 Facial Nerve Function

Facial nerve function represents a group of fundamental functions performed by the facial nerve - the seventh paired cranial nerve or simply CN VII (Gupta et al., 2013). It mainly consists of: 1) Motor functions, supplying the muscles of facial expression, the posterior belly of the digastric, the stylohyoid and the stapedius muscles with motor fibres. 2) Sensory functions, providing special taste sensation

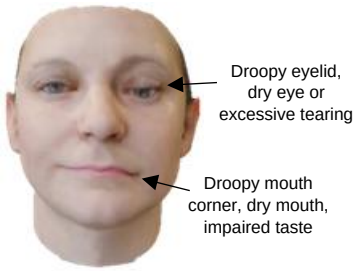


Figure 5.1: Typical symptoms of facial palsy.

from the anterior 2/3 of the tongue and general sensation from a small area around the concha of the auricle. 3) Parasympathetic functions, innervating a portion of head and neck glands, including submandibular and sublingual salivary glands. Since the facial nerve is principally composed of motor fibres, facial nerve function generally refers to the motor function manifested by various facial muscle movements, thus can be effectively evaluated from outer facial features without any obtrusive physical intervention.

5.3.1 Relationship with Facial Palsy

Once the facial nerve is damaged, the aforementioned functions will be partially or completely lost, hence causing paralysis to the affected side of the face, which is also known as facial palsy. Typical symptoms of facial palsy (see Fig. 5.1) are inability to frown, reduced elevation of the eyebrow and closure of the eye, loss of blinking and squinting control, droopy lower eyelid, decreased tearing, dropping of the mouth to the affected side, inability to whistle or blow, altered taste, etc. Facial palsy patients may subsequently suffer from various sequelae (Peitersen, 2002), including hyperkinesis, synkinesis and atrophy. All of these conditions could result in marked facial disfigurement, interrupt basic human function such as eating, drinking and speaking. The functional disability or impairment may further lead to a wide range of psychosocial problems (Bradbury et al., 2006; Fu et al., 2011; VanSWEARINGEN et al., 1998; Walker et al., 2012). Investigations carried out in Japan, UK and USA show that only the annual incidence of Bell's palsy (a typical facial palsy type) is 20 to 30 per 100,000 population (Newadkar et

al., 2016; Peitersen, 2002). It thus calls for immediate and effective action to understand and alleviate the suffering of such a large group of affected people, in which the primary step is to perform an accurate and efficient facial nerve function assessment which is a prerequisite for facial palsy diagnosis and therapy (Heaton et al., 2013; Jayatilake et al., 2013).

5.3.2 Assessment with Facial Nerve Grading Scales

To date, clinical facial nerve function assessment still relies on clinician to subjectively evaluate features such as resting symmetry, symmetry of voluntary movement and synkinesis. Targeting at providing a more uniform and accurate method for assessing facial nerve function, a variety of facial nerve grading scales such as House-Brackmann (House, 1983), Sunnybrook (Ross et al., 1996), Yanagihara (Yanagihara, 1977), FNGS 2.0 (Committee et al., 2009) and eFACE (Banks et al., 2015) have been developed over the years. These scales divide the degree of facial nerve damage into a series of discrete levels based on some rigorously-validated measures, including facial symmetry at rest, differential voluntary facial muscle movement, and secondary features such as synkinesis. Clinicians summarized the ideal characteristics of a facial nerve grading scale with current technologies: 1) perform regional scoring of facial nerve function; 2) conduct static and dynamic measures; 3) assess secondary sequelae such as synkinesis; 4) generate reproducible results with low interobserver and intraobserver variability; 5) sensitive enough to track changes over time and following interventions; 6) convenient for clinical use. A 2015 systematic review (A. Y. Fattah et al., 2015) found only Sunnybrook (see Table 5.1) fulfilled all criteria among previous grading scales.

Although sophisticated grading scales (Banks et al., 2015) are being developed for clinical applications and the discussion (Niziol et al., 2015) over the clinical effectiveness of the scales continues, all these grading scales are limited by the subjective nature of clinician-based assessment and have inherent problems such as labor-intensive, time-consuming and might yield low reproducible results with

Table 5.1: Sunnybrook grading scale.

Measure Description			Score
Resting Symmetry (compared to normal side)	0 – normal, 1 – narrow, wide or eyelid surgery	Eye	
	0 – normal, 2 – absent, 1 – less pronounced or more pronounced	Cheek (naso-labial fold)	
	0 – normal, 1 – corner dropped or corner pulled up/out	Mouth	
Symmetry of Voluntary Movement (degree of muscle excursion compared to normal side)	1 – unable to initiate movement/no movement 2 – initiated slight movement 3 – initiated movement with mild excursion 4 – movement almost complete 5 – movement complete	Forehead Wrinkle	
		Gentle Eye Closure	
		Open Mouth Smile	
		Snarl	
		Lip Pucker	
Synkinesis (degree of involuntary muscle contraction)	0 – none: no synkinesis or mass movement 1 – mild: slight synkinesis 2 – moderate: obvious but not disfiguring synkinesis 3 – severe: disfiguring synkinesis/gross mass movement of several muscles	Forehead Wrinkle	
		Gentle Eye Closure	
		Open Mouth Smile	
		Snarl	
		Lip Pucker	
Resting Symmetry Score (RSS) = score(eye, cheek, mouth) x 5 Voluntary Movement Score (VMS) = score(facial expressions) x 4 Synkinesis Score (SS) = score(facial expressions)		Composite Score = VMS - RSS - SS	

interobserver and intraobserver variability (A. Y. Fattah et al., 2015; Niziol et al., 2015). As an alternative, automated instruments enabling cost-effective, efficient, objective and quantitative facial nerve function assessment from ubiquitous visual face capture are invaluable and highly expected.

5.4 Automated Assessment from Visual Face Capture

As mentioned above, most facial nerve dysfunction is visually observable with clear static or dynamic facial signs, which motivated a lot of studies on automated facial nerve function assessment from biomedical visual capture of the face. A typical paradigm of such an instrument is illustrated in Fig. 5.2. It first uses an ordinary camera to take pictures of the patient's face when it is at rest or performing specified facial expressions. Then, computational techniques (Guarin et al., 2018; Z. Guo et al., 2017; Hsu et al., 2018; T. Wang et al., 2016) in various areas such as computer vision, image processing and machine learning are employed to objectively and quantitatively assess the facial nerve function within a certain feature space. The resulting solution can significantly reduce the subjective bias in assessment and would be easily ported to ubiquitous mobile

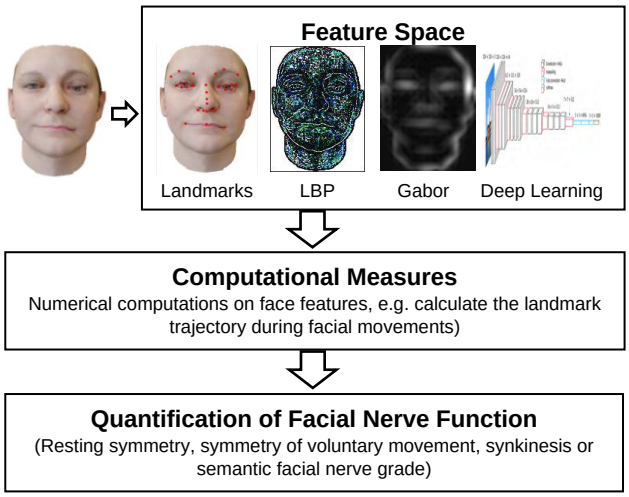


Figure 5.2: Pipeline of the automated facial nerve function assessment system.

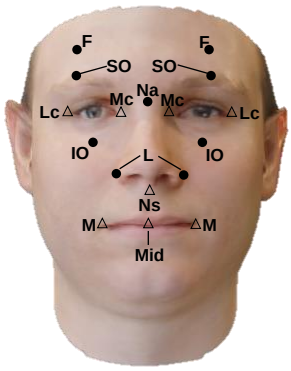
devices such as smartphones and tablets, hence has promising applicability in facial palsy diagnosis and therapy. In the following, I will systematically review the principal studies in this important area along two main dimensions – computational measures and assessment outcomes. According to the modality of the input data, computational measures can be further divided into 2D measures and 3D measures.

5.4.1 Computational Measures in 2D

Numerous computational measures on facial palsy images have been developed. They are all based on clinical measurement of facial nerve function, mainly including evaluation of facial symmetry at rest, facial movements and secondary deficits such as synkinesis (Kleiss et al., 2013). Two fundamental categories of computational measures are static measures (Hadlock & Urban, 2012; Song et al., 2017) and dynamic measures (Guarin et al., 2018; He et al., 2009), whereby facial resting symmetry and muscle movements are principally evaluated.

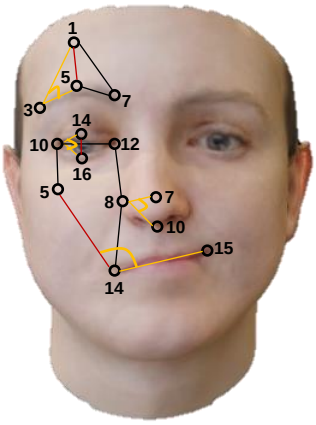
1) The Role of Facial Landmarks

A large portion of computational measures are built on top of a group of facial fiducial points called landmarks to quantify facial symmetry and movements. The



SO - on the most lateral portion of the orbital rim, above the pupil;
F - 2cm superior to point SO;
IO - placed in the most inferior fold of orbital skin, directly below the pupil;
M - the corner of the mouth;
Mid - the midline opposite the nasal spine at the center of the mouth;
Na – Nasion;
Lc - Lateral canthus;
Mc - Medical canthus.

Figure 5.3: Facial landmarks applied in (Burres, 1985).



Distance (red)
{1, 5} - Forehead distance;
{14, 16} - Palpebral distance;
{5, 14} - Cheek distance.

Area (black)
{1, 3, 5, 7} - Forehead area;
{10, 12, 5, 8, 14} - Cheek area;

Angle (yellow)
{1, 3, 5} - Forehead angle;
{10, 14, 16} - Palpebral angle;
{7, 8, 10} - Nasolabial angle;
{5, 14, 15} - Smile angle.

Figure 5.4: Typical distance, angle and area among landmarks (Hontanilla & Aubá, 2008).

pioneer work of Burres (Burres, 1985) calculated 13 distances among ten landmarks on faces at rest and during expressions to evaluate facial motor function. The points (see Fig. 5.3) were manually marked on the face with a grease pencil, and the distance was gauged with a hand-held caliper. This inefficient process was then significantly improved by applying the reflective marker (Jorge Jr et al., 2012; Linstrom, 2002; Miyazaki et al., 2000), image-editing software (Bray et al., 2010; Hadlock & Urban, 2012), image processing (Barbosa et al., 2016; Dong et al., 2008) and computer vision techniques (Guarin et al., 2018; Z. Guo et al., 2017; T. Wang et al., 2014, 2016) to automate landmark placement and distance calculation on a digitized face photograph. Evaluation of the angle and area among landmarks is also incorporated to augment the facial function quantification (see Fig. 5.4)

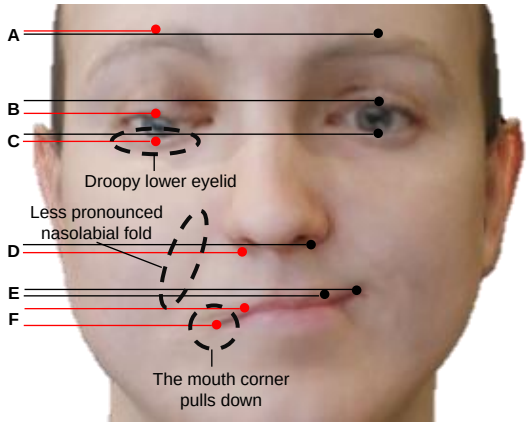


Figure 5.5: Landmark position deviations for measuring the resting asymmetry (Hadlock & Urban, 2012). Red (paralyzed side) and black (normal side) dots represent landmarks on top edge of eyebrow in mid-pupillary line (MPL), margin of upper eyelid in MPL, margin of lower eyelid in MPL, alar base, mid-upper lip position, and oral commissure position. Horizontal black and red lines indicate height of these landmarks. The vertical lines represent facial midline (based on bisection of the inter-pupillary line) (black) and the actual center of the philtrum (red). A - Resting brow ptosis, B - Superior eyelid malposition, C - Inferior eyelid malposition, D - Nasal base ptosis, E – Mid-upper lip ptosis, F - Oral commissure malposition, G - Philtrum deviation.

(Bray et al., 2010; Dulguerov et al., 2003). This initiates a basic measurement which is sensitive to facial abnormalities of spatial (topological) nature and has been widely applied in automated facial nerve function assessment. Whereas various facial landmarks have been proposed in subsequent studies (Barbosa et al., 2016; Bray et al., 2010; Dong et al., 2008; Dulguerov et al., 2003; Guarin et al., 2018; Z. Guo et al., 2017; T. Wang et al., 2014, 2016), there is a simple rule: the landmarks of interest are located close to the facial area responsible for facial movements, or on anatomical points. For example, as shown in Fig. 5.3, SO (eyebrow) and IO (lower lid) for eye closure, M and Mid for smile, Lc is on lateral canthus and Mc is on medical canthus.

2) Static Measures

The resting asymmetry is a result of muscle weakness on one side of the face. Typical symptoms (see Fig. 5.5) are droopy lower eyebrow and lower eyelid, the

mouth corner droops, and the depth or orientation of the nasolabial fold alters. Most of these features can be effectively quantified with vertical deviations of landmark positions compared against the normal side of the face, e.g. brow ptosis, superior eyelid malposition, inferior eyelid malposition, nasal base ptosis, mid-upper lip malposition, oral commissure malposition, and philtrum deviation toward the healthy side (Hadlock & Urban, 2012). Fig. 5.5 demonstrates such deviations with paired red and black lines on a paralyzed face in repose. Difference between landmark-based triangle areas (Dulguerov et al., 2003) and angle degrees (Bray et al., 2010) from the two sides of the face is also frequently used in quantifying the asymmetry. All these measures are initially represented in image pixels, which could be further scaled by the inter-pupillary distance (the average human iris diameter is 11.77mm (Rüfer et al., 2005)) to allow "real-life" millimetric measurements on the image.

However, landmarks can hardly depict the contrast between the nasolabial folds in two sides of the face, which exhibits non-pronounced variations in topology. To address this problem, measures upon image pixel intensities could be adopted, e.g. distances between pixel intensities (G. Cheng et al., 2010; S. Wang et al., 2004; S. Wang & Qi, 2006) or mediate visual texture descriptors such as Local Binary Pattern (LBP) histogram features (He et al., 2009) and circular Gabor features (Ngo, Seo, et al., 2016) from two sides of the face.

3) Dynamic Measures

Evaluation of facial movements evoked by voluntary muscle contraction lays the basis of almost every facial nerve function assessment instrument. Despite the nearly limitless ways in which humans may move the muscles of facial expression, typical attempted movements critical in facial function and communication are frequently evaluated: forehead wrinkle, eye closure, nose wrinkle, smile and lip pucker (Burres, 1986; Dulguerov et al., 2003; Hadlock & Urban, 2012) (see Fig. 5.6).

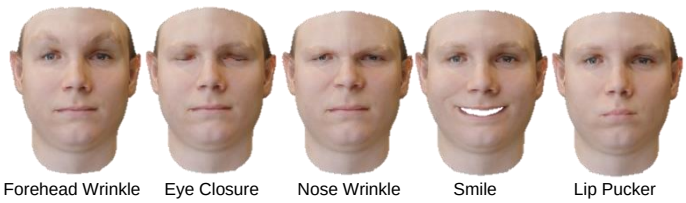


Figure 5.6: Typical facial expressions involved in evaluation of voluntary movement.

Photographs of the face in repose and with facial expressions, or videos of the face performing facial expressions are normally required and analyzed. Similar with the resting symmetry, the symmetry of voluntary muscle movement can also be efficiently measured using landmarks. Generally, changes lying in difference between landmark-based line distances (Burres, 1986; Hadlock & Urban, 2012; Kim et al., 2015) (or triangle areas (Azoulay et al., 2014; Z. Guo et al., 2017), angle degrees (Bray et al., 2010; Guarin et al., 2018)) introduced in static measures between rest and maximum movements are first gauged to quantify the muscle excursion. Then, the symmetry of voluntary movement is denoted as deviations between quantified muscle excursions on two sides of the face. Secondary defects such as synkinesis resulting from abnormal activation of muscles during expression could be measured in the same way as those mentioned above (Kleiss et al., 2013). Instead of just factoring in two states (neutral and peak) during facial movement, a few measures (Gerós et al., 2016; Miyazaki et al., 2000) are based on the trajectories (position over time) of facial landmarks. They can not only appraise abnormalities of spatial nature, but also assess temporal characteristics such as the velocity and moving direction.

As discussed in static measures, an inherent deficiency of landmark-driven measures is that they are insensitive to abnormality with obscure topological features such as changes in the nasolabial fold. The deficiency will be magnified in voluntary movement symmetry evaluation. As the zygomaticus muscle contracts during smiling, the nasolabial fold commonly deepens a lot on the normal side, while it keeps almost unchanged on the affected side. This contrast manifests clear evidence of asymmetry, thus should not be excluded. To this end, analogous

measures (He et al., 2009; McGrenary et al., 2005; Meier-Gallati et al., 1998; Ngô et al., 2016; Ngo, Seo, et al., 2016; T. Wang et al., 2014) driven by pixel intensities as those used in static measures could be applied. A simple solution is to perform a subtraction between images obtained at rest and during facial movement, then compare the luminance changes of a specific paretic area with that of the healthy side (McGrenary et al., 2005; Meier-Gallati et al., 1998). Such kind of methods however is sensitive to illumination changes, which is restricted to environment with controlled lighting. To cross this constraint, some studies resorted to robust visual texture descriptors (He et al., 2009; Ngô et al., 2016; Ngo, Seo, et al., 2016; T. Wang et al., 2014). He et al. (He et al., 2009) employed the multi-resolution LBP (MLBP) on temporal-spatial domain to extract the motion features from each region of the face. Then they assessed the symmetry of facial motion by the Resistor-Average Distance (RAD, a distance measure between two probability distributions that is closely related to the Kullback-Leibler distance) between MLBP features. NGO et al. further extended the facial texture analysis from spatial domain to frequency domain by using Gabor filters (Ngô et al., 2016), circular Gabor filters (Ngo, Seo, et al., 2016). More recent studies (Hsu et al., 2018; Sajid et al., 2018) turned to deep learning methods such as convolutional neural networks (CNNs) which have revolutionized the visual imagery analysis to extract high-level features from the face image. The extracted features are supposed to embed the most prominent image patterns probably including the facial abnormality into a compact numerical vector. A major concern about this method is that deep learning always requires a huge amount of data (typically more than 10K images) for training. Creating such a large-scale dataset is extremely expensive and time-consuming, let alone it might involve intractable ethics problems as the data exposes the privacy of patients.

As facial movements are driven by muscles located in specific facial areas (this does not apply to synkinesis which is a kind of abnormal muscle activation), regional analysis is important in measuring facial motions. For example, smiling only accounts for facial movement around the mouth region. Therefore, it is

beneficial to restrict computational measures within facial regions that are responsible for the corresponding facial movements (Barbosa et al., 2016; G. Cheng et al., 2010; Sawai et al., 2012; T. Wang et al., 2014). The face can be divided into regions according to facial landmarks (T. Wang et al., 2014) or using other image segmentation techniques (Barbosa et al., 2016; G. Cheng et al., 2010).

5.4.2 Computational Measures in 3D

An inherent shortcoming of 2D measures is that they cannot deal with out-of-plane facial movement due to the anatomical nature of skull. Gross et al. (M. M. Gross et al., 1996) found that 2D analysis underestimates 3D facial motion amplitudes by up to 43%. Mendes et al. (Mendes et al., 2014) measured the cornea surface on a 3D eyeball model created with a CAD (computer-assisted design) software, which was identified to be far more accurate than calculating only the 2D distance between the two eyelids for corneal exposure measurement. 3D analysis is hence crucial for more accurate assessment of complex facial function.

1) Landmark-based Measures

Many existing 3D measures (Gaber et al., 2015; Gerós et al., 2016; Hontanilla & Aubá, 2008; Katsumi et al., 2015; Mehta et al., 2008; Ngo, Chen, et al., 2016; Tzou et al., 2012; Vinokurov et al., 2015) are built upon the analysis of 3D facial landmark's trajectory during standardized facial movements. Distances, angles and surface between 3D landmarks on the normal side of the face are typically calculated and compared with that on the paralyzed side (Hontanilla & Aubá, 2008). During this procedure, a 3D motion capture system is employed to reconstruct and track the 3D facial geometry. Such systems were usually established with a multi-view camera setup (Hontanilla & Aubá, 2008) (see Fig. 5.7) or a mirror structure (Tzou et al., 2012). These systems required a tedious calibration process and invasive reflective markers attached on the subject's face to track 3D facial landmarks. Mehta et al. (Mehta et al., 2008) applied a different system called 3D VAS which was calibration free and was able to track a dense

Figure 5.7: Multi-camera setup (Hontanilla & Aubá, 2008), RGB-D cameras (Gaber et al., 2015; Vinokurov et al., 2015) and 3D hand-held scanner (Özsoy et al., 2019) used in 3D facial motion capture systems.

3D shape in real-time. However, the 3D VAS required color fringe patterns to be projected on the face during motion capture. It either didn't provide an efficient means to track facial landmarks which had to be manually annotated frame by frame. A few recent studies (Gaber et al., 2015; Gerós et al., 2016; Katsumi et al., 2015; Vinokurov et al., 2015) developed more compact and cost-effective 3D motion capture systems which only comprised a portable RGB-D camera (see Fig. 5.7). Meanwhile, advanced computer vision facial tracking algorithms were incorporated to further automate the 3D capture system (Gerós et al., 2016; Tzou et al., 2012; Vinokurov et al., 2015).

2) Surface-based Measures

The landmark trajectory only outlines the facial movement in a coarse manner, so it is unable to depict more in-depth morphological changes in facial soft tissue. To solve this problem, a few studies (Gibelli et al., 2018; Özsoy et al., 2019; Patel et al., 2015; Sforza et al., 2018; Taylor et al., 2014) introduced 3D surface-based measures. They first applied commercial 3D scanners such as 3dMDflex^{TM} to repetitively capture the detailed 3D geometry of the face with repose and during facial expressions in a specified period of time. Then, measures such as point-to-point root mean square (RMS) between the registered 3D point clouds of the normal side and the paralysed side, the neutral face and the morphed face to quantify face symmetry and the intensity of the facial expression (see Fig. 5.8).

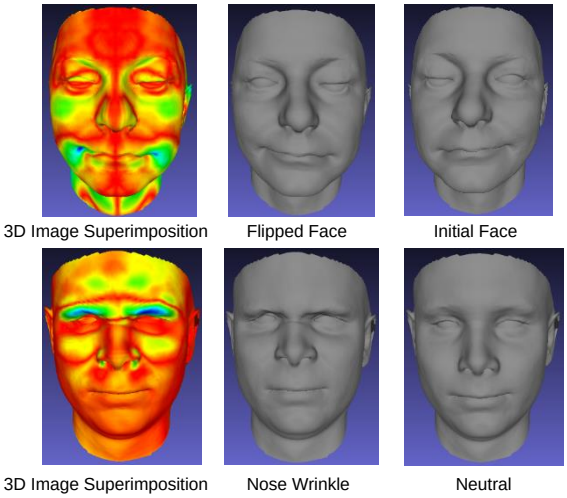


Figure 5.8: Point-to-point distance between 3D images of face and flipped face, facial expression and neutral face, for constructing 3D surface-based measures. Please note that the distance value increases from red to blue.

Statistical analysis such as ANOVA and t-test showed high intra-observer and inter-observer reproducibility of such surface-based measures, which implied a potential more reliable and accurate method to assess facial nerve function. Recent studies (Özsoy et al., 2019) found that using mobile hand-held 3D scanners, e.g. Artec^{TM} Eva (see Fig. 5.7), could achieve similar measure accuracy as that using stand-by immobile 3D scanners. This indicates that 3D surface-based measures could probably be widely applied in clinic without the need of a complicated laboratory setup.

5.4.3 Assessment Outcomes

Although a variety of automated facial nerve function assessment solutions have been proposed, their outcomes fall into two main categories: i) non-semantic numerical values quantifying static, dynamic and synkinetic facial features; ii) semantic grade of facial nerve function designed by the clinician.

The majority of existing solutions belongs to the first category, which output at least one aforementioned computational measure in high precision. For instance, as reported in (Hontanilla & Aubá, 2008), results from a 3D measurement

instrument called FACIAL CLIMA varied from the caliper results an average of 0.11% regarding the distance measured and 0.41% regarding the angles measured. The intra-rater (test-retest) reliability of these measurements is quite high, with an intra-rater correlation greater than 0.9 (Bray et al., 2010). Most of these solutions however stays at the method discussion phase, only a few of them (Bray et al., 2010; Gerós et al., 2016; Guarin et al., 2018; Hadlock & Urban, 2012; Tzou et al., 2012) have been implemented into prototypes. As presented in (Gerós et al., 2016), a typical system of its kind embeds the facial function measuring algorithms into a user-friendly graphical interface to acquire and process facial motion data. The analysis outcomes are organized into a graph named facegram to present the measures with plots and tables. Tools such as pointers, zooming and line axis tracings are provided to facilitate the user interaction. Whilst these solutions provide detailed insights and quantifications about abnormal conditions, they still need clinicians to judge the severity of facial nerve dysfunction.

Solutions in the second category instead aim to quantify the facial nerve function according to a specific facial nerve grading scale designed by the clinicians. To achieve this target, machine learning techniques should be applied to build a predictive model which is trained on labelled data and capable of making predictions on new data. The model is called classifier when the prediction is of assigning an unseen data sample into one or more predefined classes, or regressor if the prediction output is continuous. When applying to my case, the data refers to images of facial movement from either a healthy subject or a facial palsy patient, and the prediction is the grade of facial nerve function. If the grade is discrete, a classifier is employed, otherwise a regressor is employed. The classifier or regressor is trained on a group of facial movement data (from both healthy subject and facial palsy patient) which has been graded by clinicians, using methods such as support vector machine (SVM) (Azoulay et al., 2014; Z. Guo et al., 2017; Kim et al., 2015; T. Wang et al., 2016), artificial neural network (ANN) (McGrenary et al., 2005), k-nearest-neighbor (KNN) (He et al., 2009) or hybrid classifier (Barbosa et al., 2016). For a new subject, the solution first extracts computational

features from his/her facial movement data, then calls a pre-trained classifier to map the features to the facial nerve function grade defined in the grading scale. The most frequently used grading scale is House-Brackmann scale (HBS) which divides the facial nerve function into six levels (Z. Guo et al., 2017; He et al., 2009; Sajid et al., 2018; T. Wang et al., 2016), followed by Yanagihara scale (YGS) (Ngô et al., 2016; Ngo, Seo, et al., 2016) and Sunnybrook scale (SGS, see Table 5.1) (Azoulay et al., 2014). The grade could also simply be a binary value indicating whether the subject has facial palsy or not (Barbosa et al., 2016; Kim et al., 2015), or if a specific face region is paralyzed (Hsu et al., 2018). The reported classification accuracy (by comparing the predicted grade against that from the clinician) varies a lot, ranging from 49.9% (Z. Guo et al., 2017) to 95.5% (Azoulay et al., 2014). As the dataset used for evaluation and the grading scale applied are different in studies, it's difficult to compare solutions from each other. Additionally, although many studies (Azoulay et al., 2014; Kim et al., 2015) claim that their solutions have been implemented into a computer program or mobile application, only one presents the system prototype (O'Reilly et al., 2010).

5.5 Discussion

Although a number of automated facial nerve function assessment instruments have been developed, none of them has gained widespread use in clinical practice to date. The reliability of these instruments lacks sufficient clinical validation, which is the major concern. The instrument's inadequate applicability also remains a big obstacle for it to become widely accessible. According to outcomes discussed above, existing instruments can be broken down into two types – non-semantic instrument (nsINST) and semantic instrument (sINST). nsINST targets at supplying the clinician with objective quantification of facial nerve function. sINST is built on top of a clinical grading scale, which requires a specialized model training on some clinician-labelled data. In the following, I will discuss the

limitations of both instruments respectively and envisage the future directions in this field.

5.5.1 Limitations of Existing nsINST

Despite the capability of providing high-precision facial function measures comparable against calipers (Hontanilla & Aubá, 2008), the clinical effectiveness of nsINST remains the primary question as it lacks thorough and rigorous clinical validation. Researchers or clinical practitioners are consistently working on this issue. Bray et al. (Bray et al., 2010) tested their SMILE system (for measuring lip excursion during smiling with face photographs) on a database of 20 free gracilis transfer procedures with subjectively excellent results and follow-up of 4 to 12 months following single-stage surgery or 12 to 18 months following second-stage surgery to evaluate outcomes in facial reanimation. In (Tzou et al., 2012), Tzou et al. reported 241 facial palsy patients were filmed and analyzed with their 3D facial motion capture system, accounting for more than 1,000 videos made to track the rehabilitation progress after each operational therapy. These tests validate the reliability of nsINST to some extent, however the sample size and variety involved in the cohort study yet seem to be insufficient for a medical tool. The instruments hence fail to gain a wider (e.g. international) agreement and are only locally accepted.

Another essential problem is that existing nsINST are highly constrained by ineffective motion capture techniques used. To ensure the accuracy of measures, during motion capture, the patient's head was often required to stay in a stable position relative to the camera (Z. Guo et al., 2017). Intrusive markers were normally required to be placed on the face to track facial landmarks (Gerós et al., 2016; Tzou et al., 2012). These not only cause discomfort or annoyance to the patient, but also prevent the system from being applied in a broader range of circumstances such as the patient's home.

Table 5.2: Comparison of automated facial nerve grading systems and criteria.

Methods	Static Features	Dynamic Features	3D Features	Deep Features	Grading Scale*	Dataset	Prediction Accuracy
Hsu et al. 2018	×	✓	×	✓	BFR	Hsu et al. 2018	93%
Sajid et al. 2018	×	✓	×	✓	HBS	Sajid et al. 2018	92.9%
Z. Guo et al. 2018	✓	✓	×	×	HBS	Z. Guo et al. 2018	49.9%
Ngo, Chen, et al. 2016	✓	✓	✓	×	YGS	Kihara et al. 2011	66.5%
Ngo, Seo, et al. 2016	✓	✓	×	×	YGS	Kihara et al. 2011	81.2%
T. Wang et al. 2016	✓	✓	×	×	HBS	T. Wang et al. 2014	89.9%
Azoulay et al. 2014	✓	✓	×	×	B	Azoulay et al. 2014	95.5%
He et al. 2009	✓	✓	×	×	HBS	He et al. 2009	69.3%

*Grading scale: B – a binary value indicating if the subject has facial palsy or not; BFR – a binary value indicating if a specific face region is paralyzed or not.

5.5.2 Limitations of Existing sINST

sINST utilizes machine intelligence to grade facial nerve function according to a semantic facial nerve grading scale. However, current sINSTs are still far from satisfying clinical requirements and have apparent limitations. As the performance of a sINST relies on the grading scale applied, the extracted features in the prediction process and the dataset for training, the following discussion will concentrate on these three aspects.

Since a sINST is built on top of a facial nerve grading scale, its reliability highly depends on the robustness of the applied scale. As described in the previous section, clinicians have specified several characteristics for an ideal facial nerve grading scale and find only Sunnybrook meets all the criteria (A. Y. Fattah et al., 2015). However, Table 5.2 shows that most existing sINSTs were built upon less advanced grading scales such as HBS and YGS, which divide the overall facial nerve function into a few discrete levels with only general explanations. The potential effect of such sINSTs is therefore limited. The reason that previous sINSTs preferred to use less sophisticated grading scales is supposed to have two folds: i) Sophisticated grading scales such as Sunnybrook require accurate sub-scores for different facial regions and facial expressions, which is more arduous

for the clinician to grade. This makes the training data more expensive to acquire. ii) Modelling the grade consisting of semantic sub-grades will introduce more complexity to the machine learning algorithm. Therefore, to develop a sINST, it is important to find a good trade-off between the grading scale's robustness and the machine learning model's complexity.

An ideal feature is supposed to contain critical information of facial nerve function, mainly including resting symmetry, symmetry of voluntary facial movement and synkinesis. As introduced in the previous section, these features can be acquired from static, dynamic and 3D measures. I thus summarize the representative sINST according to the measures they performed. As shown in Table 5.2, almost all sINST conducted static and dynamic measures. In (Hsu et al., 2018) and (Sajid et al., 2018), deep learning methods were applied to extract high-level features that output a promising prediction accuracy rate. Meanwhile, rare sINST utilized 3D measures. As discussed in Section 5.4, 3D measures have shown to be superior against 2D measures, hence should cause more attention. It can also be noticed that the prediction accuracy rates of sINST vary a lot, from 49.9% to 95.5%. Since datasets (see Table 5.3), facial nerve function grading scales and evaluation protocols (e.g. what were input to the instrument, images or videos? How many samples were for training and testing?) adopted in these sINST are different from each other, the accuracy value actually cannot fully reflect the instrument's true performance.

As shown in Table 5.3, datasets applied in studies are different from each other. The biggest concern is that the subject cohort involved in existing datasets seems to be insufficient. For example, in (Z. Guo et al., 2017), most HBS grades contain less than 5 subjects. Meanwhile, most datasets (Z. Guo et al., 2017; Kihara et al., 2011; T. Wang et al., 2016) only include subjects from an identical ethnic background. Their applicability to other ethnic groups needs to be further verified. Another issue is none of these datasets is publicly accessible, causing no benchmark available to develop a widely accepted sINST and further push it to the clinical use.

Table 5.3: Datasets used to develop sINSTs.

Dataset	Descriptions
Hsu et al. 2018	**Source**: collected from YouTube. **Data**: 32 videos of 21 facial palsy patients. **Label**: Paralyzed face region – eyes/mouth in a video frame was outlined with an average rectangle plotted by three specialists.
Sajid et al. 2018	**Source:** collected from UCSD, PCDS and online resources**. Data:** 2, 000 real faial palsy images and 5,000 synthetic facial palsy images generated by GANs (Goodfellow et al., 2014). **Label:** each image was labelled with a HBS score.
Z. Guo et al. 2018	**Source:** captured from recruited subjects. **Data:** 480 images (480 × 640 pixels) selected from 160 facial expression videos captured from 32 subjects (14 males, 18 females). Each subject performed 5 expressions - expressionless, raising eyebrows, closing eyes, bulging cheek and showing teeth. 3 images randomly selected from each video. **Label:** subjects were graded according to HBS - 5 in I (healthy), 2 in II, 5 in III, 4 in IV, 5 in V and 11 in VI.
T. Wang et al. 2014	**Source:** captured from recruited subjects. **Data:** 570 facial epression images from 57 facial palsy patients (31 females, 26 males). 2 images per patient for each of 5 facial expressions – raising eyebrows, closing eyes, screwing up nose, plumping cheek and opening mouth. **Label:** each subject was graded with a HBS score.
Azoulay et al. 2014	**Source:** captured from recruited subjects. **Data:** videos of 9 facial expressions (face at rest, strong eye elosure, weak eye closure, rasied eyebrows, closed mouth smile, big smile, puckering of lips, puff-up cheeks and stretching down lower lip) were recorded from 14 patients and 31 healthy subjects (15 females, 30 males). **Label:** three otolaryngologists independently graded the patients's facial palsy according to HBS, YGS and SGS.
Kihara et al. 2011*	**Source:** captured from recruited subjects. **Data:** multiview face images captured from 83 subjects (74 patients, 9 healthy subjects) with a multi-camera setup (7 cameras). Each subject performed 10 expressions. Each camera took 60 images (2,112 × 2,816 pixels) for each expression. **Label:** each expression was graded with a YGS score.
He et al. 2009	**Source:** captured from recruited subjects. **Data:** 197 videos (720 × 576 pixels, 500-700 frames per video) taken from subjects with Bell's palsy, trauma to the nerve from skull fracture and surgical damage, and normal subjects. Each video presents 5 facial movements. **Label:** each video was graded with HBS by a clinician.

* The dataset was proposed in (Kihara et al., 2011) and then applied in (Ngo, Chen, et al., 2016) and (Ngo, Seo, et al., 2016). However, some key information of the database reported in the three papers are inconsistent, including the number of subjects involved, which is 5 in (Kihara et al., 2011), 83 in (Ngo, Chen, et al., 2016) and 85 in (Ngo, Seo, et al., 2016). As common authors are found in all the three papers, it seems that the database has been extended after it was first reported. I reported here the version with the most details.

5.5.3 Prospect

Overall, for both nsINST and sINST, a widely acceptable benchmark database for evaluation is urgently needed. The constraint is mainly due to the high complexity and expense of data collection which could be alleviated by more extensive collaborations among practitioners across the world. It is worth pointing out that, in (Sajid et al., 2018), the authors proposed to augment the original training dataset by automatically synthesizing facial palsy images (see Fig. 5.9) with a cutting-edge deep learning method – GANs (Goodfellow et al., 2014), which is cost-effective and highly efficient. Although the synthetic facial palsy images in (Sajid et al., 2018) still need significant improvements, it inspires me to introduce a novel theory – Parallel Vision (K. Wang et al., 2017) to solve the data problem.

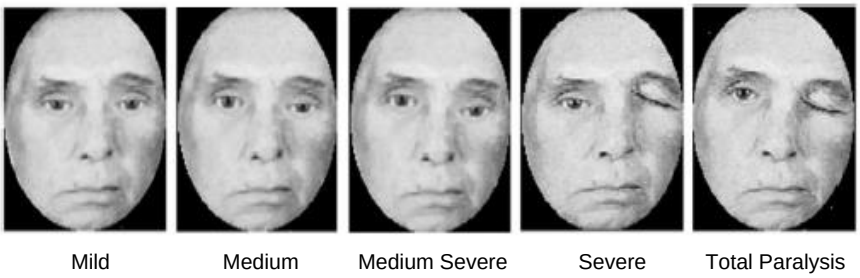


Figure 5.9: Facial palsy images synthesized in (Sajid et al., 2018) with various severity level.

Figure 5.10: Facial expressions synthesized in (Nagano et al., 2018).

Parallel Vision emphasizes the importance of photorealistic image synthesis in addressing the problems of visual perception and understanding. It comprises three stages: i) building artificial (virtual) scenes by synthesizing diverse photo-realistic data samples to simulate natural physical scenes that occur in real life; ii) conducting computational experiments on the pre-built artificial data to develop vision models (algorithms); ii) executing the vision model on the artificial data and real data concurrently to realize virtual/real interaction. Consequently, the vision model could be continuously optimized. The theory has been successfully applied in many facial analysis tasks, e.g. monocular 3D face reconstruction (Richardson et al., 2016), facial expression synthesis (Gecer et al., 2018), 3D gaze estimation (Lu et al., 2016) and facial expression recognition (Y. Wang et al., 2017). In (Y. Wang et al., 2017), the authors trained their facial expression recognition model on a dataset consisting of synthetic face images rendered from 3D facial scans and

real images captured from movies, and achieved a very promising recognition rate which outperformed the state-of-the-art by an average of 11.13% for seven basic facial expressions. Meanwhile, with only a single face image in arbitrary poses, existing face synthesis techniques can generate natural-looking face images (Nagano et al., 2018) even for those with extreme facial expressions such as asymmetric facial expressions (see Fig. 5.10). This provides solid technical supports for synthesizing photorealistic facial palsy images. I therefore believe that the Parallel Vision theory has a huge potential to fill the data gap discussed in this review and worth to be further investigated.

For a more flexible and portable assessment instrument, monocular face capture approaches (C. Cao et al., 2015; Gou et al., 2017; Thies, Zollhofer, et al., 2016; Yamaguchi et al., 2018) such as those developed in the previous three chapters are also highly needed. With only a commodity RGB camera, they can effectively capture 2D/3D facial shape and movements. As mentioned above, those information are critical for assessing facial nerve functions through metrics such as resting symmetry, symmetry of voluntary movement and synkinesis. It therefore opens up a great opportunity for applying the proposed face capture approaches. What's more, the main end-users of the assessment instruments are facial palsy patients whose faces suffer from various degrees of paralysis (please refer to Section 5.3.1 for more information). Typically, the paralyzed face is asymmetric no matter it's in repose or with an expression, hence is very hard to be accurately captured. This poses a big challenge to the capture approach, but will in turn motivate the development of the approach's performance to a new level.

As a future work, I propose to utilize the developed face capture approaches to address the problem of automated facial nerve function assessment with the following four steps: 1) Collect facial expression images from at least 100 facial palsy patients. Label each image with sparse landmarks and a clinical grade. Obtain each patient's 3D facial geometry and texture with a commercial 3D scanner. 2) Train 2D/3D facial tracking models on the collected data using the methods proposed in Chapter 2 and Chapter 3. 3) Learn a regression from the

tracked 2D/3D face to the clinical grade. 4) Build a sINST by incorporating the learned regression, and validate its performance in practical facial palsy management applications. I believe the envisaged work can be accomplished with a profound interdisciplinary collaboration with clinicians and researchers in related areas.

5.6 Conclusion

Effective and objective assessment of facial nerve function in facial palsy patients is essential to gauge severity of disease, monitor progression over time, evaluate the outcomes of therapeutic interventions and facilitate communications among practitioners, however still remains unresolved. Automated instrument working on biomedical visual face capture utilizes image processing, computer vision and machine learning techniques to carry out computational measures on facial nerve function in a highly efficient and widely accessible way, is appearing as a promising solution. By reviewing the principal studies related to this topic, this study found that though many automated instruments have been developed, they are still at a preliminary stage far from meeting clinical requirements. These instruments are severely limited by the lack of a rigorously validated benchmark database and insufficient incorporation of advancements in other areas such as monocular 3D face tracking and deep learning. To eliminate these obstacles, broader and deeper interregional and interdisciplinary collaborations are necessary and highly anticipated. Advancements in computer vision and deep learning areas such as the Parallel Vision theory (K. Wang et al., 2017), unconstrained monocular 2D/3D face reconstruction and tracking techniques (C. Cao et al., 2015; Gou et al., 2017; Thies, Zollhofer, et al., 2016; Yamaguchi et al., 2018) should be incorporated much more to further develop the instrument. In particular, this introduces a good application avenue for the face capture approaches developed in the previous chapters. Those approaches can accurately capture 2D/3D facial shape and motion which are critical in assessing facial nerve functions. On the

other side, the new application scenario will in turn push the approaches to tackle challenging cases such as capturing asymmetric facial expressions. Based on these insights, the chapter envisages a potential pathway to apply the proposed face capture approaches onto automated facial nerve function assessment.

Chapter 6
Summary and Outlook

This thesis developed novel facial performance capture approaches with low data and computational cost. The approaches are applicable to different performance modalities and use-cases. Specifically, they can capture the face from sparse 2D facial landmarks to dense 3D facial geometry, from traditional visual scenes via a monocular RGB camera to novel virtual reality (VR) scenarios that provide immersive interaction via a head-mounted display (HMD) with integrated EMG sensors. Whereas these approaches were developed for capturing different aspects of facial performance or for different application scenarios, they are connected closely with each other in terms of the algorithms applied. For example, the 2D and 3D facial tracking methods proposed in Chapter 2 and Chapter 3 are both built upon the cascaded regression method. The 3D face reconstruction algorithms adopted in Chapter 3 and Chapter 4 are both optimization-based and utilize the energy term of 2D facial landmark alignment in the objective function. This enables a smooth transition between the works of different chapters and makes the whole research more articulated and integrated. The thesis also explored a new application avenue of the developed face capture solutions, which targets at automated facial nerve function assessment for facial palsy management. The detailed contributions of this thesis are elaborated as follows:

Chapter 2 proposed a novel optimization subspace learning method to improve the canonical Supervised Descent Method - SDM (Xiong & De la Torre, 2013) for more robust 2D face alignment and tracking. The proposed method is named as Multi-subspace SDM (MS-SDM). It divides the original intricate optimization space of face alignment into multiple simpler subspaces using k-means on facial appearance features. Within each subspace, a generic descent map (or shape regression) that is able to move the initial facial shape towards the ground-truth

shape can be learned more easily via SDM. Given an unseen face image, MS-SDM first employs a Naive Bayes classifier to predict the subspace label, then calls the corresponding subspace shape regression to detect facial landmarks. The experimental results on challenging face datasets showed that MS-SDM can detect landmarks more accurately than SDM. By further developing a mobile facial tracking application, I demonstrated the potential of applying MS-SDM for 2D facial tracking on live video streams.

Chapter 3 extended the facial performance capture to more challenging 3D facial tracking from a monocular RGB camera. It developed a new boosting method called globally-optimized modular boosted ferns (GoMBF) to solve multi-modal facial motion regression via compositional learning. GoMBF is a deep composition of several boosted ferns with each was initially trained for predicting partial motion parameters of the same modality and later refined towards the whole regression target with a global optimization step. It shows stronger predictive power and a fast learning speed in comparison with the conventional boosted ferns (X. Cao et al., 2014). By cascading a sequence of GoMBFs (GoMBF-Cascade) for regressing facial motion parameters, I achieved the state-of-the-art 3D facial tracking performance even using a small training set. It thus provides a highly elegant and practical 3D facial tracking solution to real-world applications. This chapter further deeply investigated the effect of synthetic face images on training GoMBF-Cascade. I synthesized three types of face images with different naturalness levels for training, and compared the regression models trained on real data, on synthetic data and on mixed data. In my experiments, the GoMBF-Cascade models trained purely on synthesized images showed poor tracking performance on real videos and became more biased after incorporating the synthetic data into training. These two insights can benefit a variety of non-deep learning face image analysis tasks where the labelled real data is hard to obtain.

Chapter 4 addressed the intractable problem of capturing the VR HMD user's facial expression for immersive face-to-face communication/interaction in virtual environment. It proposed to integrate lightweight EMG sensors into the HMD in

an unobtrusive manner to capture the user's facial movements which are significantly occluded by the HMD, and map the captured movements to a user-specific 3D face model to recover the user's facial expression with high-fidelity. The proposed method is an innovative combination of a classic monocular 3D face reconstruction algorithm (Thies, Zollhofer, et al., 2016) and a pioneering facial biosensing technique – Faceteq (Mavridou et al., 2017). It extends the face capture from the traditional visual scene to the novel VR context, which paves the way to many new and exciting VR applications.

Chapter 5 deepened my research by exploring a new direction for applying the developed facial performance capture approaches to solve real-world problems. It identified a novel application avenue – automated facial nerve function assessment from visual face capture, which is crucial for facial palsy management. In the chapter, I systematically reviewed the most relevant and representative studies on related topics, identified the principal challenges and indicated several promising directions for future work. To the best of my knowledge, this is the first study of its kind to be reported so far. I believe that this review can not only inspire researchers in the field of face capture, but also is helpful to clinical practitioners, neurologists and bioengineers.

As shown above, this thesis addressed a number of existing problems in the field of facial performance capture. It in turn posed new questions and inspired new ideas which are worth for further research:

i) It can be found that the MS-SDM method proposed in Chapter 2 has not utilized the complementary information between different subspaces. This indicates a promising direction to further improve the MS-SDM's performance: combining subspace-specific shape regressions via compositional learning. The underlying issue resembles the one of integrating the motion parameter–specific boosted ferns into a GoMBF as studied in Chapter 3. With a similar global optimization step as that used in Chapter 3, an improved shape regression is supposed to be learned for more accurate facial landmark detection.

ii) Quality face images with ground-truth labels are essential to train a robust face capture model. However, collecting such data is normally time-consuming and labor-intensive. This is a long-standing bottleneck that has severely restricted the development of face capture solutions, especially in the field of 3D face capture as discussed in Chapter 3. As an alternative, synthesizing face images for training is highly economic and efficient. It has been successfully applied in some face analysis tasks, such as 3D face reconstruction (Y. Guo, Zhang, Cai, Jiang, et al., 2018; Richardson et al., 2016) and face frontalization (Y. Wang et al., 2017). With the advent of more powerful generative adversarial networks - GANs (Nagano et al., 2018; Saito et al., 2017), face images with high-resolution, extremely photorealistic facial texture and background can be synthesized now. The synthesis process can even be manipulated by tuning some specific facial attributes such as head pose, facial expression and illumination by using a conditional GAN (Tewari et al., 2020). This stimulates a promising research direction - using GANs to generate high-quality face images for training the face capture model.

iii) As discussed in Chapter 4, capturing the VR HMD user's facial performance with high-fidelity is of vital importance to achieve more immersive communication in virtual environment, however has not been well-studied. Whereas this thesis developed a practical capture solution to this largely unexplored problem, there is still huge space for further improvement. For example, more efficient EMG signal processing techniques could be developed to achieve real-time performance, a significant-sized database covering a wide range of AUs and the corresponding EMG signals could be created for training to enable continuous facial expression capture, or utilizing more lightweight and sensitive infrared cameras that can be attached onto the headset ergonomically to visually capture the HMD wearer's facial performance.

iv) I found that existing face capture outcomes have mainly served for content creation or human-machine interaction in the fields of entertainment, social media and security, while have rarely been applied in facial biometrics for medical and healthcare purpose. On the other side, Chapter 5 revealed that the cutting-edge

facial performance capture techniques could play an important role in automated facial nerve function assessment for facial palsy management, but had never been fully utilized. To this end, it is timely and imperative to deeply incorporate the advanced face capture techniques into assessing facial nerve function. I believe that such an interdisciplinary integration could push the facial palsy management which currently still relies on the clinician judgement to a new level, while creating more application opportunities for various face capture solutions such as those proposed in Chapter 2 to Chapter 4.

References

Aldrian, O., & Smith, W. A. (2012). Inverse Rendering of Faces with a 3D Morphable Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(5), 1080–1093.

Azoulay, O., Ater, Y., Gersi, L., Glassner, Y., Bryt, O., & Halperin, D. (2014). Mobile Application for Diagnosis of Facial Palsy. *Proc. 2nd Int. Conf. Mobile Inf. Technol. Med.*

Banks, C. A., Bhama, P. K., Park, J., Hadlock, C. R., & Hadlock, T. A. (2015). Clinician-graded electronic facial paralysis assessment: the eFACE. *Plastic and Reconstructive Surgery*, *136*(2), 223e-230e.

Barbosa, J., Lee, K., Lee, S., Lodhi, B., Cho, J. G., Seo, W. K., & Kang, J. H. (2016). Efficient quantitative assessment of facial paralysis using iris segmentation and active contour-based key points detection with hybrid classifier. *BMC Medical Imaging*, *16*(1), 23.

Basri, R., & Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(2), 218–233.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2930–2940.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proc. The 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., & Zafeiriou, S. (2017). 3D face morphable models "in-the-wild." *Proc. IEEE CVPR*, 5464–5473.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018). Large scale 3d morphable models. *International Journal of Computer Vision*, *126*(2–4), 233–254.

Bradbury, E. T., Simons, W., & Sanders, R. (2006). Psychological and social factors in reconstructive surgery for hemi-facial palsy. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *59*(3), 272–278.

Bray, D., Henstrom, D. K., Cheney, M. L., & Hadlock, T. A. (2010). Assessing outcomes in facial reanimation: evaluation and validation of the SMILE system for measuring lip excursion during smiling. *Archives of Facial Plastic Surgery*, *12*(5), 352–354.

Burres, S. A. (1985). Facial biomechanics: the standards of normal. *The Laryngoscope*, *95*(6), 708–714.

Burres, S. A. (1986). Objective grading of facial paralysis. *Annals of Otology, Rhinology & Laryngology*, *95*(3), 238–241.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Cao, C., Bradley, D., Zhou, K., & Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, *34*(4), 1–9.

Cao, C., Hou, Q., & Zhou, K. (2014). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics*, *33*(4), 1–10.

Cao, C., Weng, Y., Lin, S., & Zhou, K. (2013). 3D shape regression for real-time facial animation. *ACM Transactions on Graphics*, *32*(4), 1–10.

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*(3), 413–425.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, *107*(2), 177–190.

Cha, J., Kim, J., & Kim, S. (2016). An IR-based facial expression tracking sensor for head-mounted displays. *Proc. IEEE SENSORS*, 1–3.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Chen, H., Li, J., Zhang, F., Li, Y., & Wang, H. (2015). 3D model-based continuous emotion recognition. *Proc. IEEE CVPR*, 1836–1845.

Cheng, G., Dong, J., Wang, S., & Qu, H. (2010). Evaluation of facial paralysis degree based on regions. *Proc. IEEE International Conference on Knowledge Discovery and Data Mining*, 514–517.

Cheng, S., Kotsia, I., Pantic, M., & Zafeiriou, S. (2018). 4DFAB: A Large Scale 4D Database for Facial Expression Analysis and Biometric Applications. *Proc. IEEE CVPR*, 5117–5126.

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., & Zafeiriou, S. (2018). A comprehensive performance evaluation of deformable face tracking “in-the-wild.” *International Journal of Computer Vision*, *126*(2–4), 198–232.

Cohn, J. F., & Schmidt, K. (2003). The timing of facial motion in posed and spontaneous smiles. *Proc. Active Media Technology*, 57–69.

Committee, F. N. D., Vrabec, J. T., Backous, D. D., Djalilian, H. R., Gidley, P. W., Leonetti, J. P., Marzo, S. J., Morrison, D., Ng, M., Ramsey, M. J., & Schaitkin, B. M. (2009). Facial nerve grading system 2.0. *Otolaryngology—Head and Neck Surgery*, *140*(4), 445–450.

Cristinacce, D., & Cootes, T. F. (2006). Feature detection and tracking with constrained local models. *Proc. BMVC*, 3.

Dementhon, D. F., & Davis, L. S. (1995). Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, *15*(1–2), 123–141.

Dollár, P., Welinder, P., & Perona, P. (2010). Cascaded pose regression. *Proc. IEEE CVPR*, 1078–1085.

Dong, J., Ma, L., Li, Q., Wang, S., Liu, L. A., Lin, Y., & Jian, M. (2008). An approach for quantitative evaluation of the degree of facial paralysis based on salient point detection. *Proc. International Symposium on Intelligent Information Technology Application Workshops*, 483–486.

Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, *111*(15), E1454–E1462.

Dulguerov, P., Wang, D., Perneger, T. V., Marchal, F., & Lehmann, W. (2003). Videomimicography: the standards of normal revised. *Archives of Otolaryngology–Head & Neck Surgery*, *129*(9), 960–965.

Ekman, P. (2002). Facial action coding system (FACS). *A Human Face*.

Ekman, P., & Rosenberg, E. I. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *Proc. IEEE CVPR*, 5562–5570.

Fattah, A., Borschel, G. H., Manktelow, R. T., Bezuhly, M., & Zuker, R. M. (2012). Facial palsy and reconstruction. *Plastic and Reconstructive Surgery*, *129*(2), 340e-352e.

Fattah, A. Y., Gurusinghe, A. D., Gavilan, J., Hadlock, T. A., Marcus, J. R., Marres, H., Nduka, C. C., Slattery, W. H., & Snyder-Warwick, A. K. (2015). Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity. *Plastic and Reconstructive Surgery*, *135*(2), 569–579.

Fu, L., Bundy, C., & Sadiq, S. A. (2011). Psychological distress in people with disfigurement from facial palsy. *Eye*, *25*(10), 1322–1326.

Gaber, A., Taher, M. F., & Wahed, M. A. (2015). Quantifying facial paralysis using the kinect v2. *Proc. IEEE EMBC*, 2497–2501.

Gecer, B., Bhattarai, B., Kittler, J., & Kim, T. K. (2018). Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model. *Proc. ECCV*, 217–234.

Gerós, A., Horta, R., & Aguiar, P. (2016). Facegram–Objective quantitative analysis in facial reconstructive surgery. *Journal of Biomedical Informatics*, *61*, 1–9.

Gibelli, D., Codari, M., Pucciarelli, V., Dolci, C., & Sforza, C. (2018). A quantitative assessment of lip movements in different facial expressions through 3-dimensional on 3-dimensional superimposition: a cross-sectional study. *Journal of Oral and Maxillofacial Surgery*, *76*(7), 1532–1538.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Proc. Advances in Neural Information Processing Systems*, 2672–2680.

Gou, C., Wu, Y., Wang, F. Y., & Ji, Q. (2017). Coupled cascade regression for simultaneous facial landmark detection and head pose estimation. *Proc. IEEE ICIP*, 2906–2910.

Gross, M. M., Trotman, C. A., & Moffatt, K. S. (1996). A comparison of three-dimensional and two-dimensional analyses of facial motion. *The Angle Orthodontist*, *66*(3), 189–194.

Gross, R., Matthews, I., Cohn, J., Kanade, T., & Baker, S. (2010). Multi-pie. *Image and Vision Computing*, *28*(5), 807–813.

Gruebler, A., & Suzuki, K. (2014). Design of a wearable device for reading positive expressions from facial emg signals. *IEEE Transactions on Affective Computing*, *5*(3), 227–237.

Guarin, D. L., Dusseldorp, J., Hadlock, T. A., & Jowett, N. (2018). A machine learning approach for automated facial measurements in facial palsy. *JAMA Facial Plastic Surgery*, *20*(4), 335–337.

Guo, Y., Zhang, J., Cai, J., Jiang, B., & Zheng, J. (2018). CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(6), 1294–1307.

Guo, Y., Zhang, J., Cai, L., Cai, J., & Zheng, J. (2018). Self-supervised CNN for Unconstrained 3D Facial Performance Capture from an RGB-D Camera. *ArXiv Preprint ArXiv:1808.05323*.

Guo, Z., Dan, G., Xiang, J., Wang, J., Yang, W., Ding, H., Deussen, O., & Zhou, Y. (2017). An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(3), 835–841.

Gupta, S., Mends, F., Hagiwara, M., Fatterpekar, G., & Roehm, P. C. (2013). Imaging the facial nerve: a contemporary review. *Radiology Research and Practice*.

Hadlock, T. A., & Urban, L. S. (2012). Toward a universal, automated facial measurement tool in facial reanimation. *Archives of Facial Plastic Surgery*, *14*(4), 277–282.

Halaki, M., & Ginn, K. (2012). Normalization of EMG signals: To normalize or not to normalize and what to normalize to. *Computational Intelligence in Electromyography Analysis-a Perspective on Current Applications and Future Challenges*, 175–194.

Hamedi, M., Salleh, S. H., & Noor, A. M. (2015). Facial neuromuscular signal classification by means of least square support vector machine for MuCI. *Applied Soft Computing*, *30*, 83–93.

Hamedi, M., Salleh, S. H., & Swee, T. T. (2011). Surface electromyography-based facial expression recognition in Bi-polar configuration. *Journal of Computer Science*, *7*(9), 1407.

Hamedi, M., Salleh, S. H., Ting, C. M., Astaraki, M., & Noor, A. M. (2016). Robust facial expression recognition for MuCI: a comprehensive neuromuscular signal analysis. *IEEE Transactions on Affective Computing*, *9*(1), 102–115.

He, S., Soraghan, J. J., O'Reilly, B. F., & Xing, D. (2009). Quantitative analysis of facial paralysis using local binary patterns in biomedical videos. *IEEE Transactions on Biomedical Engineering*, *56*(7), 1864–1870.

Heaton, J. T., Knox, C. J., Malo, J. S., Kobler, J. B., & Hadlock, T. A. (2013). A system for delivering mechanical stimulation and robot-assisted therapy to the rat whisker pad during facial nerve regeneration. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *21*(6), 928–937.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hontanilla, B., & Aubá, C. (2008). Automatic three-dimensional quantitative analysis for evaluation of facial movement. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *61*(1), 18–30.

House, J. W. (1983). Facial nerve grading systems. *The Laryngoscope*, *93*(8), 1056–1069.

Hsu, G. S. J., Huang, W. F., & Kang, J. H. (2018). Hierarchical Network for Facial Palsy Detection. *Proc. IEEE CVPRW*.

Hu, L., Ma, C., Luo, L., & Li, H. (2015). Single-view hair modeling using a hairstyle database. *ACM Transactions on Graphics*, *34*(4), 1–9.

Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y. C., & Li, H. (2017). Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics*, *36*(6), 1–14.

Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W. J., Ratsch, M., & Kittler, J. (2016). A multiresolution 3d morphable face model and fitting framework. *Proc. The 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

Huynh, L., Chen, W., Saito, S., Xing, J., Nagano, K., Jones, A., Debevec, P., & Li, H. (2018). Mesoscopic Facial Geometry Inference Using Deep Neural Networks. *Proc. IEEE CVPR*, 8407–8416.

Imaging, D. (2020). *DI4D PRO System*. http://www.di4d.com/systems/di4d-pro-system/

Jayatilake, D., Isezaki, T., Teramoto, Y., Eguchi, K., & Suzuki, K. (2013). Robot assisted physiotherapy to support rehabilitation of facial paralysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(3), 644–653.

Jeni, L. A., Cohn, J. F., & Kanade, T. (2015). Dense 3D face alignment from 2D videos in real-time. *Proc. IEEE FG*, 1–8.

Jian, M., & Lam, K. M. (2015). Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*(11), 1761–1772.

Jiang, L., Zhang, J., Deng, B., Li, H., & Liu, L. (2018). 3D Face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, *27*(10), 4756–4770.

Jorge Jr, J. J., Pialarissi, P. R., Borges, G. C., Squella, S. A. F., de Gouveia, M. D. F., Saragiotto Jr, J. C., & Gonçalves, V. R. (2012). Objective computorized evaluation of normal patterns of facial muscles contraction. *Brazilian Journal of Otorhinolaryngology*, *78*(2), 41–51.

Katsumi, S., Esaki, S., Hattori, K., Yamano, K., Umezaki, T., & Murakami, S. (2015). Quantitative analysis of facial palsy using a three-dimensional facial motion measurement system. *Auris Nasus Larynx*, *42*(4), 275–283.

Kemelmacher-Shlizerman, I., & Basri, R. (2010). 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 394–405.

Kihara, Y., Duan, G., Nishida, T., Matsushiro, N., & Chen, Y. W. (2011). A dynamic facial expression database for quantitative analysis of facial paralysis. *Proc. IEEE ICCIT*, 949–952.

Kim, H. S., Kim, S. Y., Kim, Y. H., & Park, K. S. (2015). A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors*, *15*(10), 26756–26768.

Kleiss, I. J., Hohman, M. H., Quatela, O. E., Marres, H. A., & Hadlock, T. A. (2013). Computer - Assisted Assessment of Ocular Synkinesis: A Comparison of Methods. *The Laryngoscope*, *123*(4), 879–883.

Kortylewski, A., Egger, B., Morel-Forster, A., Schneider, A., Gerig, T., Blumer, C., Reyneke, C., & Vetter, T. (2018). Can Synthetic Faces Undo the Damage of Dataset Bias to Face Recognition and Facial Landmark Detection? *ArXiv Preprint ArXiv:1811.08565*.

Laine, S., Karras, T., Aila, T., Herva, A., Saito, S., Yu, R., Li, H., & Lehtinen, J. (2017). Production-level facial performance capture using deep convolutional neural

networks. *Proc. ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 1–10.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. *Proc. ECCV*, 679–692.

Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P. L., Nicholls, A., & Ma, C. (2015). Facial performance sensing head-mounted display. *ACM Transactions on Graphics*, *34*(4), 1–9.

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, *36*(6).

Linstrom, C. J. (2002). Objective facial motion analysis in patients with facial nerve dysfunction. *The Laryngoscope*, *112*(7), 1129–1147.

Liu, Q., Deng, J., & Tao, D. (2015). Dual sparse constrained cascade regression for robust face alignment. *IEEE Transactions on Image Processing*, *25*(2), 700–712.

Lou, J., Cai, X., Dong, J., & Yu, H. (2020). Real-time 3D facial tracking via cascaded compositional learning. *IEEE Transactions on Image Processing*.

Lou, J., Cai, X., Wang, Y., Yu, H., & Canavan, S. (2019). Multi-subspace supervised descent method for robust face alignment. *Multimedia Tools and Applications*, *78*(24), 35455–35469.

Lou, J., Wang, Y., Nduka, C., Hamedi, M., Mavridou, I., Wang, F. Y., & Yu, H. (2019). Realistic facial expression reconstruction for VR HMD users. *IEEE Transactions on Multimedia*, *22*(3), 730–743.

Lou, J., Yu, H., & Wang, F. Y. (2019). A review on automated facial nerve function assessment from visual face capture. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(2), 488–497.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lu, F., Gao, Y., & Chen, X. (2016). Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia*, *18*(9), 1772–1782.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *Proc. IEEE CVPRW*, 94–101.

Ma, L., & Deng, Z. (2019). Real-time hierarchical facial performance capture. *Proc. ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 1–10.

Martinez, B., Valstar, M. F., Jiang, B., & Pantic, M. (2017). Automatic Analysis of Facial Actions: A Survey. *IEEE Transactions on Affective Computing*.

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*(2), 135–164.

Mavridou, I., McGhee, J. T., Hamedi, M., Fatoorechi, M., Cleal, A., Ballaguer-Balester, E., Seiss, E., Cox, G., & Nduka, C. (2017). FACETEQ interface demo for emotion expression in VR. *Proc. IEEE Virtual Reality*, 441–442.

McDonagh, S., Klaudiny, M., Bradley, D., Beeler, T., Matthews, I., & Mitchell, K. (2016). Synthetic prior design for real-time face tracking. *Proc. IEEE 3DV*, 639–648.

McGrenary, S., O'Reilly, B. F., & Soraghan, J. J. (2005). Objective grading of facial paralysis using artificial intelligence analysis of video data. *Proc. IEEE Symposium on Computer-Based Medical Systems*, 587–592.

Mehta, R. P., Zhang, S., & Hadlock, T. A. (2008). Novel 3-D video for quantification of facial movement. *Otolaryngology—Head and Neck Surgery*, *138*(4), 468–472.

Meier-Gallati, V., Scriba, H., & Fisch, U. (1998). Objective scaling of facial nerve function based on area analysis (OSCAR). *Otolaryngology-Head and Neck Surgery*, *118*(4), 545–550.

Mendes, V. M., Lasudry, J., Vandermeeren, L., & De Fontaine, S. (2014). Computerised 3D evaluation of the functional eyelid deficit in facial palsy. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *67*(2), 178–182.

Meng, Z., Han, S., & Tong, Y. (2017). Listen to Your Face: Inferring Facial Action Units from Audio Channel. *IEEE Transactions on Affective Computing*.

Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999). XM2VTSDB: The extended M2VTS database. *Proc. Second International Conference on Audio and Video-Based Biometric Person Authentication*, 965–966.

Miyazaki, S., Ishida, A., & Komatsuzaki, A. (2000). A clinically oriented video-based system for quantification of eyelid movements. *IEEE Transactions on Biomedical Engineering*, *47*(8), 1088–1096.

Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., & Li, H. (2018). paGAN: real-time avatars using dynamic textures. *ACM Transactions on Graphics*, *37*(6), 1–12.

Newadkar, U. R., Chaudhari, L., & Khalekar, Y. K. (2016). Facial palsy, a disorder belonging to influential neurological dynasty: Review of literature. *North American Journal of Medical Sciences*, *8*(7), 263.

Ngo, T. H., Chen, Y. W., Seo, M., Matsushiro, N., & Xiong, W. (2016). Quantitative analysis of facial paralysis based on three-dimensional features. *Proc. IEEE ICIP*, 1319–1323.

Ngô, T. H., Seo, M., Matsushiro, N., & Chen, Y. W. (2016). Evaluation of facial paralysis based on spatial features of filtered images. *International Journal of Bioscience, Biochemistry and Bioinformatics*, *6*(1).

Ngo, T. H., Seo, M., Matsushiro, N., Xiong, W., & Chen, Y. W. (2016). Quantitative analysis of facial paralysis based on limited-orientation modified circular Gabor filters. *Proc. IEEE ICPR*, 349–354.

Niziol, R., Henry, F. P., Leckenby, J. I., & Grobbelaar, A. O. (2015). Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *68*(4), 447–456.

O'Reilly, B. F., Soraghan, J. J., McGrenary, S., & He, S. (2010). Objective method of assessing and presenting the House-Brackmann and regional grades of facial palsy by production of a facogram. *Otology & Neurotology*, *31*(3), 486–491.

Olszewski, K., Lim, J. J., Saito, S., & Li, H. (2016). High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics*, *35*(6), 1–14.

Oskoei, M. A., & Hu, H. (2007). Myoelectric control systems—A survey. *Biomedical Signal Processing and Control*, *2*(4), 275–294.

Özsoy, U., Sekerci, R., Hizay, A., Yildirim, Y., & Uysal, H. (2019). Assessment of reproducibility and reliability of facial expressions using 3D handheld scanner. *Journal of Cranio-Maxillofacial Surgery*, *47*(6), 895–901.

Ozuysal, M., Fua, P., & Lepetit, V. (2007). Fast keypoint recognition in ten lines of code. *Proc. IEEE CVPR*, 1–8.

P., E., & Friesen, W. V. (1978). *Manual for the facial action coding system*. Consulting Psychologists Press.

Patel, A., Islam, S. M. S., Murray, K., & Goonewardene, M. S. (2015). Facial asymmetry assessment in adults using three-dimensional surface imaging. *Progress in Orthodontics*, *16*(1), 1–9.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *Proc. Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 296–301.

Peitersen, E. (2002). Bell's palsy: the spontaneous course of 2,500 peripheral facial nerve palsies of different etiologies. *Acta Oto-Laryngologica*, *122*(7), 4–30.

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. *Proc. IEEE CVPR*, 1685–1692.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2015). Global refinement of random forest. *Proc. IEEE CVPR*, 723–730.

Rezazadeh, I. M., Firoozabadi, S. M., Hu, H., & Golpayegani, S. M. R. H. (2011). A novel human–machine interface based on recognition of multi-channel facial bioelectric signals. *Australasian Physical & Engineering Sciences in Medicine*, *34*(4), 497–513.

Richardson, E., Sela, M., & Kimmel, R. (2016). 3D face reconstruction by learning from synthetic data. *Proc. IEEE 3DV*, 460–469.

Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017). Learning detailed face reconstruction from a single image. *Proc. IEEE CVPR*, 1259–1268.

Romdhani, S., & Vetter, T. (2005). Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. *Proc. IEEE CVPR*, 986–993.

Ross, B. G., Fradet, G., & Nedzelski, J. M. (1996). Development of a sensitive clinical facial grading system. *Otolaryngology—Head and Neck Surgery*, *114*(3), 380–386.

Rüfer, F., Schröder, A., & Erb, C. (2005). White-to-white corneal diameter: normal values in healthy humans obtained with the Orbscan II topography system. *Cornea*, *24*(3), 259–261.

Saeed, A., Al-Hamadi, A., & Neumann, H. (2018). Facial point localization via neural networks in a cascade regression framework. *Multimedia Tools and Applications*, *77*(2), 2261–2283.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. *Proc. IEEE CVPRW*, 397–403.

Saito, S., Li, T., & Li, H. (2016). Real-time facial segmentation and performance capture from rgb input. *Proc. ECCV*, 244–261.

Saito, S., Wei, L., Hu, L., Nagano, K., & Li, H. (2017). Photorealistic facial texture inference using deep neural networks. *Proc. IEEE CVPR*, 5144–5153.

Sajid, M., Shafique, T., Baig, M. J. A., Riaz, I., Amin, S., & Manzoor, S. (2018). Automatic Grading of Palsy Using Asymmetrical Facial Features: A Study Complemented by New Solutions. *Symmetry*, *10*(7), 242.

Samad, M. D., Diawara, N., Bobzien, J. L., Harrington, J. W., Witherow, M. A., & Iftekharuddin, K. M. (2017). A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(2), 353–361.

Sawai, N., Hato, N., Hakuba, N., Takahashi, H., Okada, M., & Gyo, K. (2012). Objective assessment of the severity of unilateral facial palsy using OKAO Vision® facial image analysis software. *Acta Oto-Laryngologica*, *132*(9), 1013–1017.

Sela, M., Richardson, E., & Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. *Proc. IEEE ICCV*, 1576–1585.

Sengupta, S., Kanazawa, A., Castillo, C. D., & Jacobs, D. W. (2018). SfSNet: Learning Shape, Reflectance and Illuminance of Faces “in-the-wild.” *Proc. IEEE CVPR*, 6296–6305.

Sforza, C., Ulaj, E., Gibelli, D. M., Allevi, F., Pucciarelli, V., Tarabbia, F., Ciprandi, D., Orabona, G. D. A., Dolci, C., & Biglioli, F. (2018). Three-dimensional superimposition for patients with facial palsy: an innovative method for assessing the success of facial reanimation procedures. *British Journal of Oral and Maxillofacial Surgery*, *56*(1), 3–7.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. *Proc. IEEE CVPRW*, 50–58.

Song, A., Xu, G., Ding, X., Song, J., Xu, G., & Zhang, W. (2017). Assessment for facial nerve paralysis based on facial asymmetry. *Australasian Physical & Engineering Sciences in Medicine*, *40*(4), 851–860.

Sumner, R. W., & Popović, J. (2004). Deformation transfer for triangle meshes. *ACM Transactions on Graphics*, *23*(3), 399–405.

Suzuki, K., Nakamura, F., Otsuka, J., Masai, K., Itoh, Y., Sugiura, Y., & Sugimoto, M. (2017). Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. *Proc. IEEE Virtual Reality*, 177–185.

Taylor, H. O., Morrison, C. S., Linden, O., Phillips, B., Chang, J., Byrne, M. E., Sullivan, S. R., & Forrest, C. R. (2014). Quantitative facial asymmetry: using three-dimensional photogrammetry to measure baseline facial surface symmetry. *Journal of Craniofacial Surgery*, *25*(1), 124–128.

Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H. P., Pérez, P., Zollhofer, M., & Theobalt, C. (2020). StyleRig: rigging styleGAN for 3d control over portrait images. *Proc. IEEE CVPR*, 6142–6151.

Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., & Theobalt, C. (2018). Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *Proc. IEEE CVPR*, 2549–2559.

Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., & Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. *Proc. IEEE ICCVW*, 1274–1283.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. *Proc. IEEE CVPR*, 2387–2395.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *ArXiv Preprint ArXiv:1610.03151*.

Tzou, C. H. J., Pona, I., Placheta, E., Hold, A., Michaelidou, M., Artner, N., Kropatsch, W., Gerber, H., & Frey, M. (2012). Evolution of the 3-dimensional video system for

facial motion analysis: ten years' experiences and recent developments. *Annals of Plastic Surgery*, *69*(2), 173–185.

Valstar, M. F., & Pantic, M. (2006). Biologically vs. logic inspired encoding of facial actions and emotions in video. *Proc. IEEE ICME*, 325–328.

VanSWEARINGEN, J. M., Cohn, J. F., Turnbull, J., Mrzai, T., & Johnson, P. (1998). Psychological distress: linking impairment with disability in facial neuromotor disorders. *Otolaryngology—Head and Neck Surgery*, *118*(6), 790–796.

Velusamy, S., Kannan, H., Anand, B., Sharma, A., & Navathe, B. (2011). A method to infer emotions from facial action units. *Proc. IEEE ICASSP*, 2028–2031.

Vinokurov, N., Arkadir, D., Linetsky, E., Bergman, H., & Weinshall, D. (2015). Quantifying hypomimia in Parkinson patients using a depth camera. *Proc. International Symposium on Pervasive Computing Paradigms for Mental Health*, 63–71.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Walker, D. T., Hallam, M. J., Ni Mhurchadha, S., McCabe, P., & Nduka, C. (2012). The psychosocial impact of facial palsy: our experience in one hundred and twenty six patients. *Clinical Otolaryngology*, *37*(6), 474–477.

Wang, C., Shi, F., Xia, S., & Chai, J. (2016). Realtime 3D Eye Gaze Animation Using a Single RGB Camera. *ACM Transactions on Graphics*, *35*(4), 1–14.

Wang, K., Gou, C., Zheng, N., Rehg, J. M., & Wang, F. Y. (2017). Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives. *Artificial Intelligence Review*, *48*(3), 299–329.

Wang, N., Gao, X., Tao, D., Yang, H., & Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing*, *275*, 50–65.

Wang, S., Li, H., Qi, F., & Zhao, Y. (2004). Objective facial paralysis grading based on Pface and eigenflow. *Medical and Biological Engineering and Computing*, *42*(5), 598–603.

Wang, S., & Qi, F. (2006). Compute aided diagnosis of facial paralysis based on pface. *Proc. IEEE EMBC*, 4353–4356.

Wang, T., Dong, J., Sun, X., Zhang, S., & Wang, S. (2014). Automatic recognition of facial movement for paralyzed face. *Bio-Medical Materials and Engineering*, *24*(6), 2751–2760.

Wang, T., Zhang, S., Dong, J., Liu, L. A., & Yu, H. (2016). Automatic evaluation of the degree of facial nerve paralysis. *Multimedia Tools and Applications*, *75*(19), 11893–11908.

Wang, Y., Yu, H., Dong, J., Jian, M., & Liu, H. (2017). Cascade support vector regression-based facial expression-aware face frontalization. *Proc. IEEE ICIP*, 2831–2835.

Wang, Y., Yu, H., Dong, J., Stevens, B., & Liu, H. (2016). Facial expression-aware face frontalization. *Proc. ACCV*, 375–388.

Weng, C. H., Lai, Y. H., & Lai, S. H. (2016). Driver drowsiness detection via a hierarchical temporal deep belief network. *Proc. ACCV*, 117–133.

Weng, Y., Cao, C., Hou, Q., & Zhou, K. (2014). Real-time facial animation on mobile devices. *Graphical Models*, *76*(3), 172–179.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, *113*(6), 767–791.

Xia, Y., Lou, J., Dong, J., Li, G., & Yu, H. (2018). SDM-based means of gradient for eye center localization. *Proc. IEEE PiCom*, 862–867.

Xiong, X., & De la Torre, F. (2015). Global supervised descent method. *Proc. IEEE CVPR*, 2664–2673.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. *Proc. IEEE CVPR*, 532–539.

Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., & Li, H. (2018). High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics*, *37*(4), 1–14.

Yanagihara, N. (1977). Grading of facial nerve palsy. *Proc. Third International Symposium on Facial Nerve Surgery*, 533–535.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. *Proc. IEEE FG*, 211–216.

Yoon, J. S., Shiratori, T., Yu, S. I., & Park, H. S. (2019). Self-supervised adaptation of high-fidelity face models for monocular performance tracking. *Proc. IEEE CVPR*, 4601–4609.

Yu, H., Garrod, O. G., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, *36*(3), 152–162.

Yu, X., Lin, Z. L., Zhang, S., & Metaxas, D. N. (2016). Nonlinear Hierarchical Part-Based Regression for Unconstrained Face Alignment. *Proc. IJCAI*, 2711–2717.

Zhang, S., Dong, J., & Yu, H. (2017). Automatic 3D face recovery from a single frame of a RGB-D sensor. *Proc. BMVCW*.

Zhang, Y., Liu, S., Yang, X., Shi, D., & Zhang, J. J. (2016). Sign-correlation partition based on global supervised descent method for face alignment. *Proc. ACCV*, 281–295.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. *Proc. ECCV*, 94–108.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. *Proc. IEEE CVPR*, 4998–5006.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2016). Unconstrained face alignment via cascaded compositional learning. *Proc. IEEE CVPR*, 3409–3417.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3d solution. *Proc. IEEE CVPR*, 146–155.

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. *Proc. IEEE CVPR*, 787–796.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. *Proc. IEEE CVPR*, 2879–2886.

Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., & Theobalt, C. (2018). State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum*, *37*(2), 523–550.

Appendix

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information

UNIVERSITY OF PORTSMOUTH

Postgraduate Research Student (PGRS) Information			**Student ID:**	831945
PGRS Name:	Jianwen Lou			
Department:	CTS	**First Supervisor:**	Prof. Hui Yu	
Start Date: (or progression date for Prof Doc students)		**2016-11-14**		
Study Mode and Route:	Part-time ☐ Full-time ☒	MPhil ☐ PhD ☒	MD ☐ Professional Doctorate ☐	

Title of Thesis:	Facial Performance Capture from Visual Input and EMG Signals
Thesis Word Count: (excluding ancillary data)	32658

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	16/LO/1780

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:		**Date:** 31/08/2020

UPR16 – April 2018

NHS
Health Research Authority
South East Coast - Brighton & Sussex Research Ethics Committee
Health Research Authority
Ground Floor, Skipton House
80 London Road
London
SE1 6LH

Telephone: 02071048308
Fax:

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

08 February 2017

Mr Charles Nduka
Queen Victoria Hospital
Holtye Road
East Grinstead
RH19 3DZ

Dear Mr Nduka

Study title:	**Validation of the MIRROR facial expression tracking application in healthy subjects and facial paralysis patients**
REC reference:	**16/LO/1780**
IRAS project ID:	**139558**

Thank you for your letter responding to the Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this opinion letter. Should you wish to provide a substitute contact point, require further information, or wish to make a request to postpone publication, please contact hra.studyregistration@nhs.net outlining the reasons for your request.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for NHS permission for research is available in the Integrated Research Application System, www.hra.nhs.uk or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publically accessible database within 6 weeks of recruitment of the first participant (for medical device studies, within the timeline determined by the current registration and publication trees).

There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will

be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from the HRA. Guidance on where to register is provided on the HRA website.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites

NHS sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Non-NHS sites

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document	*Version*	*Date*
Contract/Study Agreement [Signature page from study agreement]	1.0	15 July 2016
Contract/Study Agreement [Study agreement with QVH]	1.0	15 July 2016
Copies of advertisement materials for research participants [Poster advertising for participants]	1.0	18 March 2016
Covering letter on headed paper [Cover letter]	1.0	18 July 2016
IRAS Application Form [IRAS_Form_21092016]		21 September 2016
Letter from sponsor [Confirmed CND as authorised signatory for Emteq]		18 August 2016
Letter from sponsor		16 September 2016
MHRA Notice of No Objection Letter (Medical Devices) and relevant correspondence [Email from MHRA confirming app is not a medical device]	1.0	08 June 2016
Other [Response Letter to REC]		21 November 2016
Other [Peer review evidence]		21 November 2016
Other [PIS - Validation Phase - Healthy Control]	1.1	21 November 2016
Other [PIS - Validation Phase - Healthy Control (Tracked Changes)]	1.1	21 November 2016
Other [PIS - Validation Phase - FP Patient]	1.1	21 November 2016
Other [PIS - Validation Phase - FP Patient (Tracked Changes)]	1.1	21 November 2016
Other [PIS - Pilot Study - Healthy Control]	1.1	21 November 2016
Other [PIS - Pilot Study - Healthy Control (Tracked Changes)]	1.1	21 November 2016
Other [PIS - Pilot Study - FP Patient]	1.1	21 November 2016
Other [PIS - Pilot Study - FP Patient (Tracked Changes)]	1.1	21 November 2016
Other [Consent - Healthy Control - 24-01-17]	1.1	24 January 2017
Other [Consent - Healthy Control - 24-01-17 TRACKED]	1.1	24 January 2017
Other [Consent - FP Patient - 24-01-17]	1.1	24 January 2017
Other [Consent - FP Patient - 24-01-17 TRACKED]	1.1	24 January 2017

Participant consent form [Participant consent form]	1.0	18 March 2016
Participant information sheet (PIS) [Patient information sheet]	1.0	18 March 2016
Research protocol or project proposal [Research protocol]	1.0	15 March 2016
Summary CV for student [Student CV]	1.0	18 July 2016
Summary CV for supervisor (student research) [CND supervisor short CV]	1	21 July 2016
Summary CV for supervisor (student research) [RMK supervisor CV]		18 August 2016

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached document *"After ethical review – guidance for researchers"* gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at
http://www.hra.nhs.uk/hra-training/

16/LO/1780	**Please quote this number on all correspondence**

With the Committee's best wishes for the success of this project.

Yours sincerely

PP

Dr Simon Walton
Chair

Email:NRESCommittee.SECoast-BrightonandSussex@nhs.net

Enclosures: "After ethical review – guidance for researchers" *[SL-AR2]*

Copy to: *Ms Sarah Dawe, Queen Victoria Hospital NHS Foundation Trust*
Ms Sarah Dawe, Queen Victoria Hospital NHS Foundation Trust