Universitat Autònoma de Barcelona

Lifelike Humans: Detailed Reconstruction of Expressive Human Faces

Gemma Rotger Moll

Universitat Autònoma
de Barcelona

Lifelike Humans:
Detailed Reconstruction of Expressive Human Faces

A dissertation submitted by **Gemma Rotger Moll** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, October 29, 2020

Co-Director	**Dr. Felipe Lumberas** Centre de Visió per Computador Dept. Ciències de la computació Universitat Autònoma de Barcelona
Co-Director	**Dr. Antonio Agudo** Institut de Robòtica i Informàtica Industrial Consejo Superior de Investigaciones Científicas Universitat Politècnica de Catalunya
Thesis committee	**Dr. Xavier Binefa Valls** Departament de Tecnologies de la Informació i les Comunicacions. Universitat Pompeu Fabra (UPF) **Dr. Petia Ivanova Radeva** Centre de Visió per Computador Department Matemàtica Aplicada i Anàlis, Universitat de Barcelona (UB) **Dr. Angel Sappa** Centre de Visió per Computador Escuela Superior Politécnica del Litoral (ESPOL)

This document was typeset by the author using LaTeX 2ε.

ISBN:

Printed by Ediciones Gráficas Rey, S.L.

Learning is the only thing the mind
never exhausts, never fears,
and never regrets.
— Leonardo da Vinci

Out of clutter, find simplicity.
From discord, find harmony.
In the middle of difficulty
lies opportunity.
— Albert Einstein

To my family, and friends.…

Abstract

Developing human-like digital characters is a challenging task since humans are used to recognizing our fellows, and find the computed generated characters inadequately humanized. To fulfill the standards of the videogame and digital film productions it is necessary to model and animate these characters the most closely to human beings. However, it is an arduous and expensive task, since many artists and specialists are required to work on a single character. Therefore, to fulfill these requirements we found an interesting option to study the automatic creation of detailed characters through inexpensive setups. In this work, we develop novel techniques to bring detailed characters by combining different aspects that stand out when developing realistic characters, skin detail, facial hairs, expressions, and microexpressions. We examine each of the mentioned areas with the aim of automatically recover each of the parts without user interaction nor training data. We study the problems for their robustness but also for the simplicity of the setup, preferring single-image with uncontrolled illumination and methods that can be easily computed with the commodity of a standard laptop. A detailed face with wrinkles and skin details is vital to develop a realistic character. In this work, we introduce our method to automatically describe facial wrinkles on the image and transfer to the recovered base face. Then we advance to facial hair recovery by resolving a fitting problem with a novel parametrization model. As of last, we develop a mapping function that allows transfer expressions and microexpressions between different meshes, which provides realistic animations to our detailed mesh. We cover all the mentioned points with the focus on key aspects as (i) how to describe skin wrinkles in a simple and straightforward manner, (ii) how to recover 3D from 2D detections, (iii) how to recover and model facial hair from 2D to 3D, (iv) how to transfer expressions between models holding both skin detail and facial hair, (v) how to perform all the described actions without training data nor user interaction. In this work, we present our proposals to solve these aspects with an efficient and simple setup. We validate our work with several datasets both synthetic and real data, prooving remarkable results even in challenging cases as occlusions as glasses, thick beards, and indeed working with different face topologies like single-eyed cyclops.

Key words: *3D Reconstruction, High-Frequency Detail, Facial Hair Recovery, Expression Transfer.*

Resumen

Desarrollar personajes digitales similares a los humanos es un reto, ya que los humanos estamos acostumbrados a reconocernos entre nosotros y a encontrar a los CGI poco humanos. Para cumplir con los estándares de las producciones de videojuegos y películas digitales, es necesario modelar y animar a estos personajes de la manera más parecida posible a los humanos. Sin embargo, es una tarea costosa, ya que se requiere a muchos artistas y especialistas trabajando en un solo personaje. Por lo tanto, para cumplir con estos requisitos, encontramos la creación automática de personajes detallados a través de setups económicos una opción interesante para estudiar. En este trabajo, desarrollamos técnicas novedosas para conseguir estos personajes combinando diferentes aspectos que se destacan al desarrollar el realismo como detalles de la piel, pelos faciales, expresiones y microexpresiones. Examinamos cada una de las áreas mencionadas con el objetivo de recuperar cada una de las partes automáticamente sin interacción del usuario ni datos para el aprendizaje. Estudiamos los problemas buscando su robustez, pero también la simplicidad de la configuración, prefiriendo soluciones basadas en una sola imagen e iluminación no controlada y cálculos que pueden obtenerse en un ordenador portátil estándar. Una cara detallada con arrugas y detalles de la piel es vital para desarrollar un personaje realista. En este trabajo, presentamos nuestro método para describir automáticamente las arrugas faciales en la imagen y transferirlas a la cara base recuperada. Luego proponemos la recuperación del vello facial resolviendo un problema de ajuste de parámetros con un nuevo modelo de vello facial parametrizable. Por último, introducimos una función de mapeo que permite transferir expresiones y microexpresiones entre diferentes mallas, lo que proporciona animaciones realistas a nuestra cara detallada. Cubrimos todos los puntos mencionados con el enfoque puesto en aspectos clave como (i) cómo describir las arrugas faciales de una manera simple y directa, (ii) cómo recuperar 3D a partir de detecciones 2D, (iii) cómo recuperar y modelar el vello facial de 2D a 3D, (iv) cómo transferir expresiones entre modelos que contienen tanto el detalle de la piel como el vello facial, (v) cómo realizar todas las acciones descritas sin datos de entrenamiento ni interacción del usuario. En este trabajo, presentamos nuestras propuestas para resolver estos aspectos con una configuración eficiente y simple. Validamos nuestro trabajo con varios conjuntos de datos, tanto sintéticos como reales, demostrando resultados notables incluso en casos desafiantes como oclusiones por gafas, barbas densas y, incluso, trabajando con diferentes topologías faciales como cíclopes de un solo ojo.

Palabras clave: *Reconstrucción 3D, Detalle de Alta Frecuencia, Vello Facial, Transferencia de Expresiones.*

Resum

El desenvolupament de personatges digitals semblants a les persones és una tasca difícil, ja que els humans estem acostumats a reconèixer-nos entre nosaltres i trobar els CGI poc humanitzats. Per complir els estàndards de les produccions de videojocs i pel·lícules digitals és necessari modelar i animar aquests personatges el més proper als éssers humans. Tanmateix, és una tasca difícil i cara, ja que requereix molts artistes i especialistes treballant en un sol personatge. Per tant, per complir aquests requisits, trobem la creació automàtica de cares detallades mitjançant setups de baix cost una opció interessant per estudiar. En aquest treball, desenvolupem tècniques noves per aconseguir cares detallades combinant diferents aspectes que destaquen a l'hora de desenvolupar personatges realistes, detalls de la pell, pèls i expressions facials i microexpressions. Examinem cadascuna de les àrees esmentades amb l'objectiu de recuperar-les automàticament sense interacció de l'usuari ni dades de aprenentatge. Estudiem els problemes buscant la seva robustesa, però també la simplicitat de la configuració, preferint solucions basades en una sola imatge amb il·luminació incontrolada i mètodes que es poden calcular fàcilment amb un ordinador portàtil estàndard. Una cara detallada amb arrugues i detalls de la pell és vital per desenvolupar un personatge realista. En aquest treball, introduïm el nostre mètode per descriure automàticament les arrugues facials de la imatge i transferir-les a la cara base recuperada. A continuació, avancem a la recuperació del cabell facial mitjançant la resolució d'un problema de parametrització amb un nou model de cabell facial. Per últim, desenvolupem una funció de mapatge que permet transferir expressions i microexpressions entre diferents malles facials, que proporciona animacions realistes a la nostra cara detallada. Cobrim tots els punts esmentats parant atenció als aspectes clau com (i) com descriure les arrugues facials d'una manera senzilla, (ii) com recuperar 3D a partir de deteccions 2D, (iii) com recuperar i modelar el cabell facial a partir de 2D a 3D, (iv) com transferir expressions entre models amb detalls de la pell i cabells facials, (v) com realitzar totes les accions descrites sense dades d'aprenentatge ni interacció de l'usuari. En aquest treball, presentem les nostres propostes per resoldre aquests aspectes amb una configuració eficient i senzilla. Validem el nostre treball amb diversos conjunts de dades tant sintètiques com reals, obtenint resultats remarcables fins i tot en casos tan difícils com oclusions per ulleres, barbes denses, inclús treballant amb diferents topologies facials com ciclops d'un sol ull.

Paraules clau: *Reconstrucció 3D, Detall d'Alta Freqüencia, Pèl Facial, Transferència d'Expresions.*

Acknowledgements

I was barely 18 when I packed all my stuff and left my island, heaven on earth, but too small for my longings. Those who know me are used to my wild decisions, so they did not get surprised the day I announced to them I was starting this Ph.D. I have committed many insanities in my lifetime but nothing compared with these past six years. There have been all kinds of stages, peace, despair, worry, growth, accomplishment, grief, and joy. I still cannot believe that these stages are close to their end. Now it is time to sit down and look back to thank all the people that have accompanied me in the good and the bad. Standing me has a lot of merits. Thank you for being next to me.

First of all, I would like to thank my supervisors, Felipe Lumbreras, for giving me this opportunity, for his support, trust, and patience. Antonio Agudo, there are no words to thank you enough for your consistent support and for making the best of me. My co-author Francesc Moreno-Noguer for guiding me so positively and always made me feel confident with my work.

These years have been long, and I have gone from one place to another more frequently than I expected, so I hope to do not forget anyone.

CVC has been my nest, where I learn to fly in my first research years. My deepest gratitude to my LAB colleagues Coen, Enric, Agnès, Joan Ramon, Joan Mas, Miguel, Anna, Carles, Marc, Raúl, and Jordi Abella. To my research group Daniel Ponsa, Angel Sappa, Robert Benavente, Xavier Poma, Edgar Riba, Jordi Salvador, and Armin Mehri for all the support. I must also thank all the CVC staff for making my life there much simple, Claire, Ana Mari, Eva, Mari, Silvia, Montse, Gigi, Mireia, Raquel, Ainoha, Alexandra, Núria, and Meritxell. Encarna Talavera, for saving me from getting buried into the administration records several times. Jordi Gonzalez for all the support, especially to cope with Ph.D. bureaucracy and matcha tea. And Joan Serrat for his guidance during my first years. This years without all the support of my partners in crime, wouldn't have been the same: Pau Rodriguez, Pau Riba, Sounak, Albert, Arnau, Carola, Arash, Francesco, Jose Luís, Laura, Yaxing, Lichao, Lu, Xialei, Yi, Kai, Fei, Lei, Hana, Oguz, Abel, Ivet, Jose Antonio, Antonio Esteban, Flavio, Marco, Flavia, Javad, Iris, Biel, Fran Molero, Fran Pérez, Xisco, Nestor, Marc, Mario, and all the Synthios. And my innermost circle: Germán, Onur, Dena, Bojana, Felipe Codevila, Paricher, Lorena, Júlia, and Jorge, thanks for being my second family.

I also would like to thank all my colleagues in IRI for making my Wednesdays hilarious, Alberto Olivares, Alejandro, Albert Pumarola, Antonio Andriella, Adrià, Gerard, Rishabh, Juan Acevo, and Karla, to my office colleague Ana, and especially to my close friends Victor Vaquero and Aleks Taranovic.

All the people that welcomed me as a new member of the Infaimon family. Carol and Toni. Alex Fernández, Anna, Sergio, Adrián, Xavi López, Xavi Poch, Raúl, Xavi Giró, Héctor Abril, Juanma, Josep, Alfredo, Josema, Meri, Alberto, Dani, Ana Paula, Víctor Iglesias, and especially to my twisted sister Emma Wuyts and my evil spawn Carles Carreté.

In the last step of my professional life, I want to thank Miguel Ángel Muñoz and Jose Ramón Vega to trust in my abilities and capacities. Also to my colleagues Antonio Cortés, Xenia, Joan, Lídia, Josep, David, Marijan, and especially to Raimon and Amadeu for being great co-workers and even better friends.

Coming to the personal side, I must infinitely thank Dani Sánchez for being the best lifetime friend one can have. Also, to Lau Mercadal to appreciate as much as I do the taste of good beer. Gaizka Echeverria, for being my favorite sports journalist and a good friend that one can trust. Juan Fran, one can talk to you about everything at any moment. Soraya Guerrero, thanks for being the best flatmate of my undergraduate period. Irene, Fionna, Maria, and Sira for our long dog walks and Friday nights beers. Natalia and Bea for all the workouts and good moments to refill our energy levels. To the best travel squad, Manuel Curado, for saving me from eating baby pigeons in Beijing and for being my friend since then. Xavi Cortés, for being the best karaoke partner, and to blow all that noise meters up. Gabriele Piantadosi, for being the best travel buddy and unconditional supporter one can have, even in the craziest moments on the other side of the world. Thank you all for being a spice in life.

Last, I cannot be prouder of my family for their unconditional love and support. My parents, Bep and Carmen, for giving me the opportunities and experiences that have made me who I am. My sister Maria and my partners in life, Aleix and Congo, for the courage and love you gave me these years without expecting anything back. Thanks a lot to the rest of my extended family for being supportive in the hardest moments, specially Tonia, for making the cruelest car drive in my life a little bit human. I can't finish without thanking and remembering some of my loved ones who I've lost during these past six years: Marianna, Irene, Conxita, Bep, and Kim. May your light shine forever.

Contents

List of Figures

List of Tables

1 Introduction

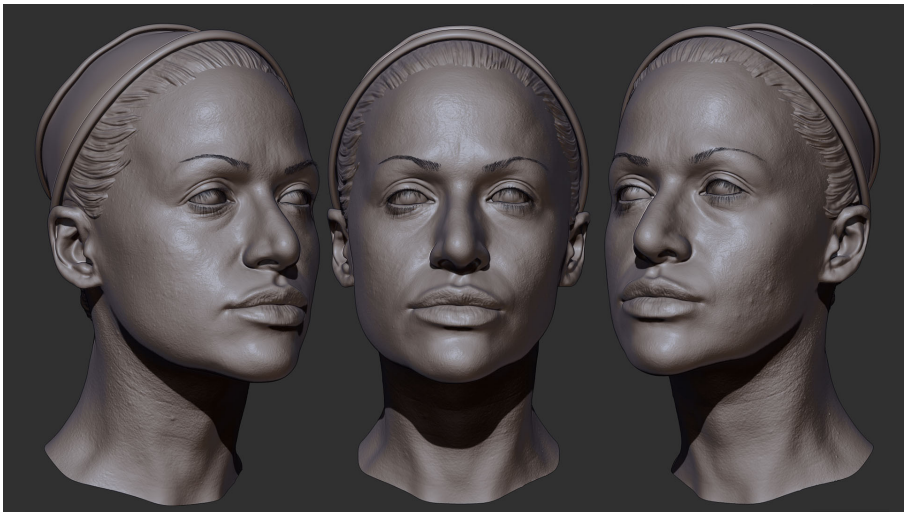


Figure 1.1 – 3D Model of Josia by Second Chance Games and Visual Effects, Inc. [58].

1.1 3D Reconstruction of Human Faces

Do we always distinguish a real person from a CGI (Computer-Generated Imagery) character? A few years ago, the answer would be simple. Indeed, at the very end, our brain is highly capable of recognizing other humans, making the generation of lifelike humans a hard task. However, in recent years, the video game and digital film industry have raised the standard of CGI characters, making them look almost human. Still, the automatic reconstruction of 3D character is far from that artist-make almost human CGIs.

Figure 1.2 – The evolution of the main character in the different videogames of the Uncharted saga. We can appreciate a large improvement between the first appearance in 2007 and the dramatic realism achieved in 2016 [71].

What we certainly recognize from these realistic digital characters is that a few specific aspects make a huge difference (see Fig. 1.1 for current solutions and Fig. 1.2 for an evolution of characters). Hence, most of the efforts in their evolution have been invested in the improvement of their textures (skin and clothing), and the naturalness of their hair. Further, it is no longer just a question of acquisition and modeling that builds an exceptional character but how it acts and interacts with the environment. Since the face is the most significant part in terms of realism, it is essential to provide the facial components with uniqueness and self-personality.

1.2 Evolution of detailed facial reconstruction

Facial reconstruction has gone through many stages since it is a topic of great interest in different areas, from computer graphics to forensics, just to name a few.. In the most demanding cases, not only the geometry needs to be estimated but also the pose, the expression, and the scene illumination. This section briefly introduces the different types of inputs that made the problem evolve during the years, making it solvable with fewer requirements.

The most convenient technique is the use of multiview stereo since most of the information can be recovered straightforward from the camera data up to manageable estimations [35, 44, 78, 87]. Photometric stereo reveals sticking results in studio conditions [28, 73]. A different approach is the use of video or monocular sequences to explore the spatio-temporal data. In this case, a priori information and

a low-rank model are expected [39, 40, 56, 70, 75]. Likewise, 3D sensors give a good prospect of 3D geometry. However, they need to be combined with an RGB sensor to recover light properties [55, 91]. The most challenging input is the single RGB image, where any depth information is known, and it cannot be computed straightforward or through a direct method. In this case, a facial morphable model [15, 21] is required, and the parameters encoding the identity and the expression are estimated through minimization functions or complex non-linear regressor functions in deep-learning models [66, 67]. This dissertation will explore and discuss in detail all of the mentioned inputs but is focused on the single image reconstruction as the most challenging input.

1.2.1 Medium and fine-level detail estimation

Detailed reconstructions have constantly been associated with multiple view techniques [35] and photometric approaches [23, 34]. Despite this, the previous methods require specific setups in controlled environments, and its results strongly depend on these conditions.

On the other side, 3D reconstruction from monocular sources is known to be an underconstrained problem [1, 2, 39, 40, 56]. In the particular case of human facial reconstruction, early proposals introduced their solutions through the use of facial 3D morphable models (3DMM) [15, 21], allowing to recover large to medium-scale geometry. Since its nature lies in a fitting problem, they benefit from a substantial dimensionality reduction of the solution space. However, there are some drawbacks to consider, and the major of them is the lack of subject-specific details such as skin wrinkles and scars. The maximum achievable details are the medium-scale features that can be explained as a linear combination of medium-scale details in the basis.

The property of RGB-D cameras capturing both color and depth had strongly benefited the reconstructions, since, most of the times, at least a coarse face can be reconstructed, depending on the approach and the environment, medium scale features can additionally be recovered [47, 54, 91, 92].

With the recent rise of deep-learning models, the first techniques retrieved 3D facial models from RGB data. From 3DMM-based models [67], to learning from synthetic data [66]. Volumetric regression was likewise applied to retrieve coarse faces from a single image [51]. Despite the exceptional results, most of the methods have a fundamental problem in their approach. These methods are unable to obtain accurate personal results since they depend on the data sources they were trained with. To obtain accurate results, they rely on refining techniques as specific detail-recovery neural networks. Moreover, it is worth noting that deep-learning approaches are data demanding. Further, the wrinkle solution space is not simple to model. For these reasons, all the mentioned methods require a postprocessing

refinement step to recover subject-specific details.

Our first contribution consists in a novel framework that joins the coarse structure of the face obtained with a volumetric neural network, with the skin wrinkles and subject-specific details modeled directly from a single RGB image with a photometric energy term. This method retrieves detailed facial geometries from the underconstrained input of a single RGB image. For convenience, we take advantage of a popular solution based on deep volumetric regressors [51] to retrieve a non-refined initialization, which has sufficient accuracy to be considered as an initial 3D mesh shape. Notwithstanding, other image-based techniques remain a valid initialization. We first estimate the scene illumination conditions required to define a photometric term. Then, we model wrinkles from 2D information with energy terms based on the photometric equation and a few smooth functions. Finally, we can precisely define wrinkles up to a depth measurement studying the relief they generate as a shading effect in the image.

1.2.2 Hair Capture

Facial hair is by far the most important feature in producing the realism, authenticity, and uniqueness of a CGI. Pessig *et al.* [63] stated that eyebrows are the most important feature for humans to recognize us. Furthermore, completing the challenge of reconstructing facial hair fibers through a single image, is a hard task due to the complex structure of the hair fibers (curl, thickness, gravity effect, self-occlusions, etc.). Additionally, the computational complexity required to solve all these unknowns is an important aspect to consider.

There have been different methods describing the hair from a single image in great detail [13, 52, 59, 61, 62, 84]. Despite that, all require specific hardware and different views. Further, except for [13], none has tried to serve for facial hair but just to generate hairstyles.

Other techniques based on a single image have managed to generate hairstyles [48, 49, 59, 88, 89], most of them are based on a low-rank hairstyle space, limiting hairstyles to linear combinations of existing ones. Moreover, none have exposed results in the generation of facial hair.

This dissertation defines a set of four energies considering several 2D hair properties to produce our 3D parametric hair model. To increase computation performance, we gather the hair fibers in separate groups depending on their 2D position, length, and orientation. For each arrangement, we capture a separate parametrization. Eyelashes and eyebrows require a particular study due to their inconsistency regarding other facial hair. The more clusters we set, the higher the computational time will be expected, but the more realism in the result. To this end, we set a final step to add realism to reduce computational complexity.

1.2.3 Blendshapes as expression models

Animation of faces and facial expression transferring has received much attention in the last years. It is a topic with a great evolution in the research for the recent years. From, the initial motion-capture methods via optical markers [42], to 2D expressions generated via neural networks [65]. First methods, were limited to determine displacements across several markers, where sparse deformations can be retrieved. Though, these methods do not support the reconstruction of small details concerning subtle gestures and micro-expressions. Latterly, dense face expression transfer was accomplished from the viewpoint of deep-learning within learning 2D-to-2D mappings from large amounts of images [9, 31]. Yet, these methods do not guarantee 2D-to-3D expression transfers.

As a final step, this dissertation presents a novel technique to transfer dense facial expressions in a 3D domain. After the detailed-face retrieval, we expose a method to densely map the recovered face to the neutral expression of a 3D low-rank template. The mapping gives the inverse transfer of expression vectors over the recovered face, and it overcomes several challenges, including mapping with different topologies, different mesh resolutions, and noisy input shapes. As the mapping is completed via local equivalent regions, it is robust to most of the geometric variations between meshes. Once the mapping is determined, we can transfer a wide range of 3D expressions to the input face as one-to-one correspondences.

1.2.4 Joint framework

The approach provides a realistic blending shapes generation from a high-resolution single image, including facial detail as well as facial hair. It is practical and compelling, and it can be effortlessly readjusted as a post-processing refinement approach for most of the contemporary single image methods. It does not require any additional data, user interaction, or training data. We expose experimental validation in an extensive assortment of synthetic and real images, including different skin properties and facial expressions, and several different hairstyles, proving the suitability of our framework to reconstruct plausible 3D detailed faces from a single image. Further, this dissertation broadly evaluates the effectiveness of our expression transfer method on an extended number of synthetic and real examples. Besides, we are considering cases with substantial topology differences between the input and the dataset shapes. For instance, we show that the expressions of our low-rank 3D dataset can be transferred to a single-eyed cyclops.

1.2.5 Applications

In this section, we detail some applications that the presented methods have now and will have in the near future.

- First of all, the methods presented in this thesis can be useful to generate new 3D synthetic data from RGB images. This application is related to popular deep-learning techniques requirement of large amounts of data.
- Considering the state-of-the-art and the rapid progress of deep-learning approaches, we believe our method can be employed to learn from synthetic data not only for 3D reconstruction purposes but in other areas as face detection, recognition, expression reenactment, etc. With the sight on increasing character realism, or transferring movement and facial deformations from motion-capture actor to CGI avatar in an automatic manner.
- Easily animate 3D characters without the need for a rigging process is a notable application for our expression transfer method since it can bring facial expressions from a motion capture actor to a human character.
- Comparing several reconstructions of the same face during the years is useful to predict the aging of a person over the following years. Our method to reconstruct detailed faces can provide data on how wrinkles evolve during the years. This information allows us to extrapolate the set up of new wrinkles for future facial ages.
- AR / VR may disrupt in the very next upcoming years, building 3D digital avatars to perform some tasks over the virtual avatar faces like try on new glasses or new hairstyles in advance would be a near-future application of our methods.

1.3 Objectives and scope

This dissertation aims to explore the facial 3D reconstruction field developing novel techniques to manage the open aspects of the problem. To this end, we focus on facial detail reconstruction as well as expression transfer with an eye on microexpressions, proposing new solutions to the detailed recovery of human faces (Part I) and effective generation of realistic blending shapes (Part II). Bearing this structure in mind, we first develop the analysis and reconstruction of detailed faces from single RGB images. In this regard, we study the wrinkle anatomy to extend the facial geometry reconstruction methods to produce further detailed results. The questions we aim to answer in this stage are:

- Is it feasible to exploit the high-frequency detail in images to create more detailed facial reconstructions?
- How can we model facial wrinkles in 3D geometries solely from visual information?
- Can we model that wrinkles up to a single (or a reduced number) of parameters?

Further, we want to examine the concept of geometric facial hair and propose a solution to approximate the geometry of realistic hairs from a single image as a particular input. We seek to answer the following questions:

- Can we model hair fibers with a reduced set of parameters?
- How can we estimate the previous parameters from 2D information?
- Do energy-based models perform well with very dense beards?
- Can we purse further realism to a 3D recovered face by generating its facial hair?

The second part of the dissertation seeks to analyze the expression transfer between different facial geometries. In this part, we propose to use a novel mapping function based on multilevel matching. It allows transferring expressions from not only 3D-to-3D but from 2D-to-3D. We answer the following questions in the course of this part:

- How to exploit facial geometry mesh points to find equivalent facial regions?
- How to adapt facial regions to be geometrically independent of the mesh resolution and facial traits?
- How can we map an expression from a human and adapt it to another human with evident facial differences?
- Can our method handle topological differences between different human faces?

This dissertation aims to bring clarity to these issues from a computer vision point of view. We aim to contribute to the 3D reconstruction process to produce more detailed results from the particular input of a single RGB image.

1.4 Outline and Contributions

This dissertation is divided into three parts. The first part deals with the aspects of producing detailed reconstructions from a single RGB image including skin wrinkles and facial hair fibers. This part starts with our proposal to retrieve skin detail in a simple and efficient manner directly from the 2D image information. In Chapter 2, we present our detailed reconstruction method for human faces. In particular, we describe a polynomial wrinkle description method and a photometric based energy minimization problem to model facial wrinkles. In this chapter, we present the following contributions:

- Code the facial wrinkles and hair spaces without any training data nor user interaction.
- Fast and accurate recovery of facial wrinkles and wounds.
- Handle detailed reconstruction under uncontrolled general lighting.

In Chapter 3, detail recovery is extended from skin to facial hair domain. In this chapter, we demonstrate our facial hair reconstruction method. On a first instance, present a 3D hair representation model based on a 3D parametric helix with different hair properties like curviness, thickness, and even a gravity-like effect. Then we turn different energy terms based on 2D characteristics to approximate the parameters of the 3D helix. Besides, we introduce a post-processing step to achieve further realism. Two remarkable contributions to this chapter are:

- Robust parametrizable model for facial hair.
- New formulation to retrieve facial hair geometry from a single RGB image.

The second part of this dissertation deals with expression transfer aspects while managing detail preservation. In Chapter 4, we depict our 2D-to-3D facial expression transfer method based on mappings between equivalent face regions, where expression vectors can map between different faces even with distinct mesh properties or facial geometries. The main goal of this part is to rig realistic characters preserving details extracted in Part 1. In this chapter, we have contributed to the following aspects:

- Transfer expressions from 2D-to-3D in a dense manner, without the use of a user-specific low-rank model.
- Transferring new expression wrinkles adapted to the source geometry.

In Chapter 5, we present a joint approach to the before-mentioned methods, where the different approaches are joined in a common setup. In this chapter, we revisit the most important aspects of each part of the pipeline and we detail con we connect the different parts to generate detailed rigger faces from a single RGB image. The final contributions presented in this chapter are:

- Holding the facial detail consistency on expression transfer.
- Animate facial hair in a simple manner, with just two parameters.

Finally, in Part 3, within Chapter 6 we conclude this dissertation discussing the relevance of the presented findings and contributions.

Details Part I

Detailed facial rig with wrinkles and facial hair, modelled by Chao Dong [33].

2 High frequency facial detail recovery from RGB images

We present a novel technique for estimating facial detail from a single RGB image. The presented approach can deal with the high-frequency detail reconstruction of human faces from solely a single image in a fast a concise manner. Our method produces reliable and detailed solutions comparable with state-of-the-art methods with much less effort and without requiring any training data. In the current chapter, we propose to formulate the detailed relief estimation as a photometry optimization problem. Joint with a simple 2D description of wrinkles and a simple 2D-to-3D conversion via energy minimization allows us to perform a fast and simple method that produces detailed faces. Additionally, our method can be employed as a refinement step for any coarse reconstruction method.

2.1 Motivation

The reconstruction of 3D human geometry has been a topic of great interest in computer vision and computer graphics in the past two decades, especially the reconstruction from RGB images and video sequences [1, 2, 22, 39, 40, 56, 70, 93]. Reconstructed human parts can be applied to many contemporary applications ranging from the movie industry with a stop at the hospital, where they can be used for medical purposes. Robotics, gaming, and human-computer interaction are also other technologies that can take advantage of human reconstructions. From all the before mentioned, a significant area of research is focused on the acquisition of human faces. Unfortunately, human faces are not easy to handle, since it is well known they are highly morphable differing widely between gender, age, ethnicity, and gesture expression, making it an uphill task. Moreover, if we add the requirement of recovering high-frequency details, it becomes a more ambitious task. For this reason, the study of high-detailed face geometry has excelled amongst other techniques.

There exist several inputs to operate. Facial reconstruction from monocular

information consists of retrieving a 3D face from a video or image sequence. It is known to be an under-constrained problem, and it requires additional priors to constrain the solution. The first monocular approaches addressed the unconstrained nature of the problem with facial 3D morphable methods (3DMM) [15, 21], which allow recovering the large-scale geometry as a combination of a low-rank identity and expression. On behalf of the parameter fitting nature of the problem, they bring in with a considerable dimensionality reduction of the solution space. However, the generality of these approaches turns them in an unsuitable alternative to retrieve subject-specific details, such as wrinkles and scars. It enforces the use of further refining alternatives to retrieve the subject-specific details.

In recent times, the use of deep-learning techniques permitted the face reconstruction from single RGB image information. With the previous background, several new proposals addressed the problem for the perspective of a single image 3DMM-based models [67], volumetric regression [51], and learning from synthetic data [66]. These techniques have proved a good performance in the reconstruction of the large-geometry of human faces. Nonetheless, the solutions still having a remarkable lack of individual detail, even training specific refining networks. A prompt solution would be to increase the amount of training data to represent the wrinkle space. However, it is not trivial in practice. Notably, most of them mentioned methods require a refinement step to retrieve individual details.

In this chapter, we introduce a novel optimization framework that combines the use of wrinkle 2D properties, directly extracted from the input image, with a photometric energy term that exploits these properties. This work aims to recover 3D faces from a single image with a high level of detail.

As an initial solution, we take advantage of contemporary solutions based on deep-learning techniques [51] to obtain a coarse initialization, which is accurate enough to estimate the initial 3D mesh shape. We selected this method for convenience, but any coarse mesh with a direct correspondence to an image can be employed. In this case, we employ deep-learning in the initialization of the coarse mesh, but no training data is required to model facial wrinkles. For the initialization of the photometric energy term, we estimate the illumination parameters with the coarse mesh. Finally, we encode wrinkles by using a set of smooth functions. In this regard, we find wrinkles can be accurately described with the extracted 2D properties from the image up to a depth parameter. We likewise find this parameter can be estimated with a photometric term since relief produced by wrinkles unfold visible shading in the image. To manage the image information, we group the connected pixels after filtering noisy measurements (removing unconnected pixels and pixels with low response values), since they have similar displacement behavior. Next, we locally fit the second-order polynomial that adequately adapts to the pixel coordinates of each wrinkle detected in the image. We compute two

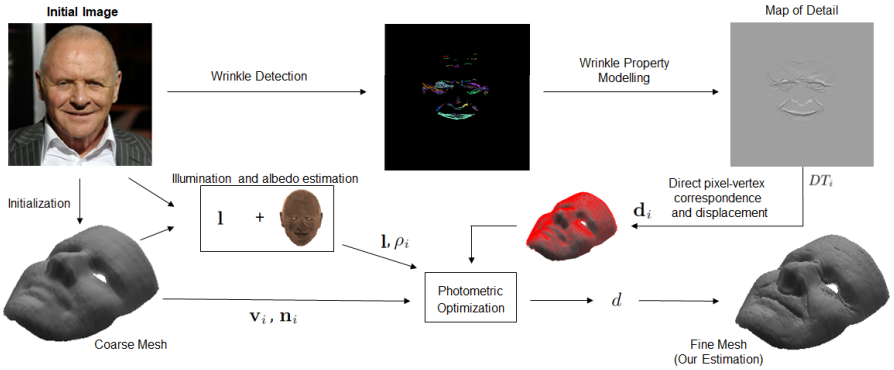


Figure 2.1 – **A graphical summary of our high-frequency detail reconstruction approach from an RGB image.** Our approach consists of a two-step method: 1) we employ a deep-learning approach to obtain a 3D coarse mesh from the image. From this geometry can be extracted the 3D locations $\mathbf{v}_i$ and normals $\mathbf{n}_i$ for the i-th point. 2) Considering the input image alone, we run a wrinkle detection algorithm based on image partial derivatives, and then model each of them. With this information, we can obtain a map of details DT_i to establish a direct correspondence between pixels in the image and vertices in the mesh, as well as the corresponding displacement vector $\mathbf{d}_i$. Finally, along with the illumination properties and albedo $(\mathbf{l}, \rho_i)$, we formulate the problem thorough a photometric optimization problem to recover the final detailed 3D mesh parametrized by depth scale d. As can be noticed, our refining approach can recover detailed areas in 3D, in contrast to initializing approaches.

distances regarding the polynomial curve. We relate it with the overall wrinkle location, orientation, height, and width to model the wrinkle displacement. Since we have a direct mapping between the image and the 3D mesh, we can directly transfer the displacements to the corresponding nodes. We have to consider, depending on the mesh and image resolutions, not all the pixels must have a correspondent vertex. As much the pixel-to-vertex correspondences lead to one-to-one, the better and more accurate the results.

Our approach offers a simple and realistic solution for high-resolution single image 3D face reconstruction, which differs from current methods since it is both fast and effective, as well as can be easily adapted as a refinement strategy for most of the initializing formulations. We present experimental validation in a wide variety of synthetic and real images, including different skin properties and facial expressions, showing the suitability of our approach even in extreme cases like

special effects facial makeup.

2.2 Related Work

Detailed facial reconstruction is the initial step and the basis of this thesis. The final result of the whole process depends on the results obtained in this step and its quality. Since facial hair reconstruction requires basis orientation. The most complex the basis is, the less uniform the hair fibers grow, and the more realism will be achieved. A similar situation occurs with expression transfer, the most detailed the basis face, the best the transferred microexpressions will fit in the facial geometry.

3D reconstruction methods can be classified in different categories depending on the required inputs. The most demanding methods (multiple-view and photometric stereo), are also the most compelling in terms of achieved results.

In recent years the most accurate results have been achieved by multiview techniques. Ghosh et al. [41] proposed a novel method to achieve detailed facial geometry with high resolution diffuse and specular photometric information from multiple viewpoints relating polarized spherical gradient illumination. Valgaerts et al. [81] stated a lightweight passive facial performance capture method able to reproduce high-quality dynamic faces from a single set of stereo cameras. Gotardo et al. [43] proposed a joint method combining Photometric Stereo, Optical Flow, and Multiview-Stereo simultaneously to achieve high-quality dynamic 3D reconstruction.

Monocular approaches take advantage of low-rank and parametric methods to represent a face solution as a linear combination of shape bases, which can be inferred directly from data or predefined in advance. However, it is well known they are highly underconstrained methods that rely on apriori information. Suwajanakorn et al. [74] presented a method to synthesize president Obama from a high-quality video. Then, with many hours of video, they can map the lip-sync from audio features to mouth shapes and textures in a photorealistic manner. Agudo et al. [2] proposed a non-rigid 3D shape model as a linear combination o mode shapes with physics-based time-varying weights computed in a temporal sliding window of a determined number of frames. Agudo et al. [6] proposed a novel shape model to encode object non-rigidity. They first recover the shape at rest and update it via spectral analysis information with physical interpretation. This analysis provides a low-rank shape basis applied to linearly expand the non-rigid deformation to the shape at rest.

RGB-D models take advantage of the depth data which gives a straightforward value of z-displacement, however, it does not take into account projective deforma-

tion and the real result is a 2.5D reconstruction if any further step is developed to achieve the 3D solution. Zollhofer et al. [91] proposed a RGBD method that required a template to reconstruct a fully rigged face. However, the shape and size variance of the face becomes a limitation to this model. In particular, when the skin produces subject-specific wrinkles and folds on aging and expression, the capture of these medium and fine-scale details compounds the problem. Li et al. [55] presented a lightweight non-parametric method to generate wrinkles for 3D facial modeling and animation. First, they build a personalized offline model with Kinect, then they explore the performance capture to obtain expression wrinkles.

Estimating the 3D reconstruction from a single image is the most under-constrained situation to this problem. Early approaches [45, 83, 86] were barely capable of producing simple objects like geometric shapes and almost flat scenes.

To estimate the facial geometry from a single image, a set of facial shapes, or a 3D Morphable Model (3DMM) [15, 21], is required. These methods bring the possibility to represent the geometry of a face as a simplified low-rank fitting problem, where regularly the identity and the expression are linearly weighted to achieve the most plausible face to fit a given image. Romdhani et al. [68] proposed a novel algorithm to estimate the 3D shape, pose, light direction and facial texture from a single RGB image by recovering the parameters of a morphable model. Huber et al. [50] developed a multi-resolution 3D Morphable Model to facilitate research in this area. Jiang et al. [53] detailed a novel method for reconstructing 3D faces from unconstrained 2D images via coarse-to-fine optimization. Employing facial landmarks for coarse reconstruction as a parametric model, and local corrective deformation fields with photometric constrain to achieve medium-detail facial shapes. Despite they achieve a significant level of detail, are time demanding since most of them rely on a coarse-to-fine strategy. Furthermore, the 3DMM may not produce a correct result if the given face is not adequately represented in the low-rank model.

As in most of the fields, deep-learning approaches attempt to provide an accurate solution to detailed face reconstruction from a single image. Richardson et al. [66] presented a semi-supervised deep framework to solve the problem of inverse rendering from a single RGB image. They use three elements in their proposal: an encoder to map the 2D input image to its representation space, a 3D decoder, and a mapping component to map 2D to 3D representation. Jackson et al. [51] proposed a direct volumetric regression neural network to obtain the 3D facial geometry from a single 2D image. Richardson en al. [67] considered the problem of recovering detailed 3D faces from single RGB inputs as a problem of data availability, so they trained a model over a large set of synthetic data to demonstrate their hypotheses. Tran et al. [80] proposed a novel framework based on nonlinear 3DMM from a large dataset of unconstrained images without their corresponding 3D scans. Their

encoder estimates the projection of both, shape and texture, and two decoders achieve the reconstruction, one for shape and the other for the texture. Chinaev et al. [30] proposed an accurate 3D facial reconstruction system via a compact and fast CNN working on realtime on mobile devices via shape regression. Nevertheless, we find in deep-learning methods the lack of a detailed reconstruction the most visible of the exposed methods, since the current deep-learning approaches are not prepared to directly predict a face with detailed facial features such as wrinkles or scars.

Our main contribution in this chapter is to develop a unified and unsupervised method that retrieves a detailed 3D mesh from a single image. It is worth noting that we do not need training data to code the wrinkle space, providing striking results even for complex shapes.

2.3 Initial 3D Face from Single Image

Before performing the detailed acquisition, it is necessary to obtain an initial 3D face estimation from a single image. It is important to remark that any method that preserves a direct vertex-pixel relation is usable. In this case, we propose to use the convolutional volumetric regression model depicted in [51]. It consists of a volumetric network that provides a volume from a single image as input. We require to adjust the network volume output to our model, which works on a triangulated mesh. We perform this conversion by retrieving the surface of the volume as our initial mesh.

The CNN resultant cropped image contains almost imperceptible details, therefore the quality of our algorithm over this image is drastically decreased. To improve the quality of the detail detection, we align the mesh with the original image at full resolution utilizing the facial features of both images extracted with [8]. This is especially important when the resultant image cancels all the facial details.

In the context of optimization, we require to estimate the scene lighting field. We take advantage of the second-order spherical harmonics [11] which can be estimated globally with a total of nine coefficients. Assuming Lambertian reflectance, we consider an initial albedo to be the average skin color over the face. The reflection model to the i-th point can be formulated as:

$$\mathbf{I}_i = \rho_i \cdot (\mathbf{l} \cdot \mathbf{Y}(\mathbf{n}_i)), \tag{2.1}$$

where $\mathbf{I}_i$ represents the image value, ρ_i is the albedo, $\mathbf{l}$ is a 1×9 vector to represent the spherical harmonic coefficients, $\mathbf{Y}$ is a 9×1 vector to designate the spherical harmonic basis, and $\mathbf{n}_i$ is the surface normal. The process to determine $\mathbf{l}$ and ρ_i is iterated until the variance is adequately low. This model is simple and fast, but it

Algorithm 1 Recovery of a detailed face from a single RGB image

Coarse initialization with [51] $\mathbf{I} \rightarrow \mathbf{V}$
Volume to mesh conversion $\mathbf{V} \rightarrow \mathbf{S}$
Scene illumination estimation $\mathbf{I}_i = \rho_i \cdot (\mathbf{l} \cdot \mathbf{Y}(\mathbf{n}_i(\mathbf{S}_i)))$
Wrinkle detection $\mathbf{I}(x_i, y_i)^k$
Detail map construction DT_i by applying Eq. (2.8)
Depth estimation d via Eq. (2.10)
Depth transferring $\mathbf{S}_i^{detailed} = \mathbf{S}_i^{coarse} + \mathbf{d}_i$

has a limitation, does not accomplish well among cast shadows.

For completness, a summary of the full pipeline is showed in Algorithm 1.

2.4 Texture analysis

Facial wrinkle modeling through a parametric formulation is a relevant area in 3D facial animation [10, 83], with applications on real-time face acquisition [18, 57]. However, the principal limitation is the parametric complexity and the requirement of a priori information to be known, such as the wrinkle properties as well as the material (skin) behavior. A common practice to describe the wrinkle in 2D is through parametric models, such as the cubic Bezier curve [10].

Unfortunately, most of the formulations require these parameters to be known, or they have to be defined by the user in advance. It limits their applicability in real scenarios. To solve this limitation, we present a simple and intuitive method to recover the detailed wrinkles that better adapt to the image plane. Our method can extract from the image all those parameters, without the need for any training data at all. Besides, our approach can operate on different mesh resolutions, and recover from large- to fine-scale details. It is worth noting that our method can adjust fine-scale wrinkles with only three points. As a limitation, a furrow with less than three associated points cannot be retrieved.

2.4.1 Localization and clustering of wrinkle pixels

To detect the wrinkle pixels on the image, we first set the images partial derivatives in both x- and y-directions. Both directions are related to determining regions of the image where variations in texture or geometry occur, i.e., enabling us to detect the wrinkles too. To avoid miss-detection from noisy measurements, we select the absolute value of measures which are higher than a particular threshold (0.25 in our experiments). Once we found the initial winkle pixel candidates, we analyze

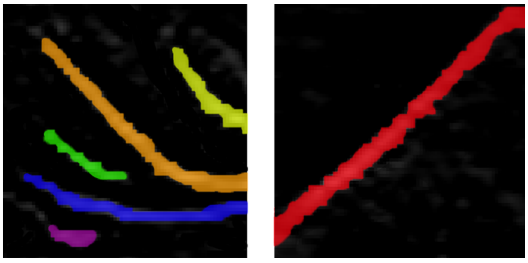


Figure 2.2 – **Wrinkle detection example.** For a given image, we identify wrinkles over the image by clustering pixels according to partial derivatives values. We obtain consistent regions with an area larger than a threshold. Each wrinkle is specified and reconstructed in an individual manner since not all the wrinkles yield the same parameterization. **Left:** We observe a detection and clustering of multiple eye lines. Each color corresponds to a separate wrinkle. **Right:** An example of a single nasolabial wrinkle is displayed, where small areas are ignored.

the connectivity between selected pixels, determining different regions where all the pixels are connected (some instances are represented in Fig.2.2). We consider K sets of 8-connected pixels where each has been previously detected a wrinkle at pixel-level. We group all the pixels in a determined wrinkle k as $\mathbf{I}(x_i^k, y_i^k)$.

2.4.2 Curve Fitting and Distance Metrics

In this section, we describe the effect of the clustered points to the geometric relief. Taking advantage of the classical curve-fitting methods for data analysis, we propose a second-order polynomial that better fits the points at each cluster. Note that for partial derivative with a major component in x, we swap axis since a well-defined function associates one, and only one, output to any particular input. Therefore, for each region k, we define the function $f(x)^k$ that best fits the data as:

$$f(x)^k = a^k x^2 + b^k x + c^k \,, \tag{2.2}$$

where the tuple (a^k, b^k, c^k) must be estimated for each cluster k. To define our parametric model over the defined furrow in a region, we propose to use the location of pixels (x_i^k, y_i^k) in the image, the curve defined by $f(x)^k$, and its corresponding centroid (x_c^k, y_c^k) such as:

$$(x_c^k, y_c^k) = \left(\frac{1}{R} \sum_i x_i^k, \frac{1}{R} \sum_i y_i^k \right) , \tag{2.3}$$

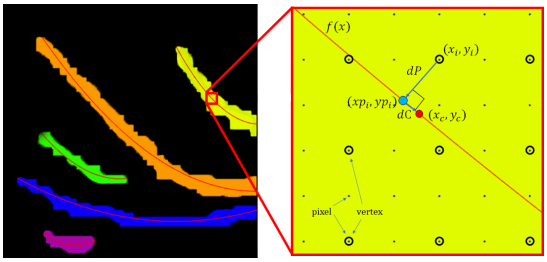


Figure 2.3 – **Curve fitting and distance metrics scheme. Left:** We can appreciate how the curves fit the given data. **Right:** Zooming view where we can appreciate the image pixel locations represented by blue dots, and these which have a corresponding mesh point by a black circle. The blue circle represents the perpendicular distance dP between a point (x_i^k, y_i^k) and $f(x)^k$, the red circle indicates the center of the wrinkle, and dC is the distance between the projected blue point (xp_i^k, yp_i^k) and the central point (x_c^k, y_c^k).

where R represents the number of pixels in the k-th region.

2.5 Wrinkle modelling

In this section, we employ the wrinkle properties previously extracted from the image. This data allows us to model wrinkles and furrows in 2D and afterward extend the model to 3D information.

To exploit the 2D properties extracted from the image, operate over all the wrinkles k where $k \in K$ represents all the wrinkles on the face geometry. Let us define the distance dP representing the perpendicular distance between a point in the wrinkle group (x_i^k, y_i^k) and the wrinkle polynomial $f(x)^k$. At the same time, we manage to compute dC as the distance between the same point projected into the curve $f(x)^k$ and the centroid of the group (x_c^k, y_c^k) (both distances are described in Fig.2.3)

Considering previous definitions, we now assume that the greatest impact of the depth parameter is where dP and dC shift to zero. The global maximum of depth variation is represented at the projection of the central point (x_c^k, y_c^k) projection. Then, to represent the smoothness in the wrinkles, it loses influence while distances dP and dC increase.

To obtain the point projection (xp_i^k, yp_i^k), we first find the tangent to $f(x)^k$ at

point x_i^k, defined as:

$$T_s = 2a^k x + b^k. \tag{2.4}$$

After that, we find the negative reciprocal slope, which is the normal slope:

$$N_s = -1/T_s, \tag{2.5}$$

N_s is also the normal line to $f(x)^k$. Finally, we compute all the possible solutions of the polynomial and select the one with a smaller distance to the point (x_i^k, y_i^k). On summary, both coordinates can be computed as follows:

$$xp_i^k = \frac{-b^k + N_s \pm \sqrt{(b^k - N_s)^2 - 4a^k(c^k + N_s x_i^k - y_i^k)}}{2a^k}, \tag{2.6}$$

$$yp_i^k = a^k xp_i^2 + b^k xp_i^k + c^k, \tag{2.7}$$

Next, we introduce the two 2D properties to model the wrinkle in the two-dimensional plane. The height h (2D) of the wrinkle it is defined by the maximum distance between any point (x_i^k, y_i^k) and the centroid (x_c^k, y_c^k) projection on $f(x)^k$. The width of the wrinkle we denote as w is the maximum distance that a point can have with $f(x)^k$ in the perpendicular direction.

According to previous descriptions, we can extend the wrinkle model to 3D by studying the relief of every point belonging to the region k according to the 2D parameters dP, dC, w, and h of every region.

As we previously described, the effect of the furrow on the center is maximum when both dP and dC are zero it softens as the distances get larger. We now take advantage of this assumption and define a map of detail DT. It represents the wrinkle depth with a normalized scale [-1, 1]. Its value for every i-th pixel in the region is defined as:

$$DT_i = \left(1 - \frac{dC_i}{h}\right) \cdot \left(-exp\left(\frac{-dP_i}{w}\right)\right). \tag{2.8}$$

It is important to normalize DT between -1 and 1. Hence, while the wrinkle sunk in the lowest values of DT, the remaining pixels on the outer part (with larger values in DT), slightly lift to produce a more realistic effect. The values not referring to any wrinkle are set as zero.

DT provides information on how the skin wrinkles along with the image, depending on the distances dP and dC. Further, it provides a smooth map of simulated wrinkles. Consequently, the smoothness quality of the resultant detailed

shape is guaranteed, considering the close pixels have a related value. However, meshes with a poor resolution may need a smoothing step to adjust the result to the mesh geometry, since the resultant detail may be excessively sharp. The most the resolution in both, the image and the mesh, the better. Moreover, as much the one-to-one correspondences between pixels and mesh vertices, smoother and better-detailed results.

The 3D is finally obtained by transferring the values from the map of detail to the mesh by direct pixel-vertex correspondence. Each vertex of the mesh $\mathbf{v}_i$ is displaced along its normal direction $\mathbf{n}_i$ according to the value of DT_i in order to determine the corresponding displacement $\mathbf{d}_i$ defined as:

$$\mathbf{d}_i = \mathbf{n}_i \cdot d \cdot DT_i \,, \tag{2.9}$$

where $d \cdot DT_i$ is a scalar representing the displacement magnitude in the direction of the normal vector.

2.5.1 Depth Estimation via Energy Optimization

With the 2D wrinkle properties, the remaining process is to retrieve the 3D displacements produced by the wrinkles on the regular reconstructed face. To this end, we benefit from the well known photometric properties and minimize the described photometric loss function.

$$\begin{aligned} &\mathscr{E}(d) = \underset{d}{\operatorname{argmin}} \sum_{i=1}^{I} \|\mathbf{I}_i - \rho_i \cdot (\mathbf{l} \cdot \mathbf{Y}(\mathbf{n}_i(\mathbf{v}_i + \mathbf{d}_i)))\|_2^2 \\ &\text{subject to} \quad \mathbf{d}_i = \mathbf{n}_i \cdot d \cdot DT_i, \end{aligned} \tag{2.10}$$

where the displacement vector $\mathbf{d}_i$, is a function of the parameter d we have to estimate. $\mathbf{n}_i(\mathbf{v}_i + \mathbf{d}_i)$ refers to the normal of the surface with the estimated displacement. DT is previously normalized between -1 and 1, thus d determines the exact scale of the shift. This optimization is solved with the Matlab least-squares solver.

2.6 Experimental Evaluation

In this section, we introduce the quantitative and qualitative evaluation of our method performing on real and synthetic data to validate our approach. Synthetic data is employed to analyze numerically our approach while real data serve to obtain results on a wide range of subjects with different gender, ethnicities, age, and facial gestures. From the results obtained in both datasets, we can ensure the competitivity of our method relating to other state-of-the-art techniques.

2.6.1 Synthetic data tests

To analyze numerically our method, we have employed a set that simulates different facial expressions with their corresponding expression wrinkles. This set es extracted from Victor dataset [39], which includes several 3D meshes on a wide variety of facial gestures: pain, angry, sing, smile, kiss, sad, or surprise to name a few. It includes plenty of medium wrinkles and a few fine detail wrinkles around the eyes, however, they are subtle and cannot be easily captured at small image resolutions.

Unfortunately, we cannot evaluate directly our approach since there is not an available one-to-one correspondence and the vertex difference between the ground truth mesh and reconstructed mesh is too large to recount (gt mesh has around 5k vertex around the full head, face and neck while the reconstructed mesh has around 10k vertex concentrated on the face only). To solve this inconvenience, and to provide a quantitative evaluation, we align both 3D meshes and then computing an error between virtual closest points on the meshes by following the normal direction. It means to find the equivalent point in the other mesh even when there is no direct equivalence. To this end, we follow the method described in the literature [69] to determine correspondences between aligned meshes with different element resolution. For additional details, we suggest the reading of this paper. Finally, we also found a comparison concerning state-of-the-art approaches.

For each expression, we evaluate the 3D error with respect to the ground truth as ϵ_{3D} as $\epsilon_{3D} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{v}(i)' - \mathbf{v}_{gt}(i)\|$, where $\mathbf{v}(i)'$ represents our 3D estimation of the i-th point and $\mathbf{v}_{gt}(i)$ is the corresponding ground truth, respectively.

We first initialize our approach to obtain the initial 3D coarse face with the CNN-based method [51] which we employed as a baseline. As we claimed beforehand, this solution is insufficient to recover medium and fine-scale detail and most of the wrinkles and facial details are missed. Even trough it is enough for initialization since our formulation can model every wrinkle in the image and transfer it to the coarse mesh, obtaining a detailed and realistic solution.

We numerically evaluate our approach on eight facial expressions (neutral, sad, angry, pain, surprise, kiss, happy, smile, and sing) and establish a comparison with respect to the baseline [51]. These results are summarized in Table2.1. As it is confirmed, our approach outperforms current solutions in terms of 3D accuracy. However the numerical improvement in the geometric error is not much striking since the distance shifted is small, we can appreciate a greater difference if we adopt the same metric to the displaced vertices only.

We additionally provide in Table2.1 the number of wrinkles detected by our algorithm. We consider this information as an indirect metric to evaluate the performance of our approach since a priori a higher number of disclosed wrinkles

provides further detailed surfaces. Still, this number should not be excessively high to avoid false wrinkles due to noisy measurements.

Expre.		Neutral	Sad	Angry	Pain	Surprise	Kiss	Happy	Sing	Average
[51]	ϵ_{3Dtot}	0.022	0.040	0.039	0.037	0.064	0.050	0.055	0.059	0.046
	ϵ_{3Dwri}	0.808	0.811	0.814	0.814	0.613	0.820	0.811	0.813	0.788
	$\mathscr{E}(d)_{tot}$	0.393	0.518	0.261	0.494	0.259	0.422	0.240	0.133	0.340
	$\mathscr{E}(d)_{wri}$	0.157	0.155	0.160	0.154	0.171	0.154	0.163	0.149	0.158
Ours	ϵ_{3Dtot}	**0.021**	**0.034**	**0.036**	**0.036**	**0.042**	**0.044**	**0.051**	**0.054**	**0.039**
	ϵ_{3Dwri}	**0.791**	**0.797**	**0.801**	**0.801**	**0.602**	**0.801**	**0.797**	**0.799**	**0.774**
	$\mathscr{E}(d)_{tot}$	0.389	0.508	0.253	0.486	0.254	0.414	0.233	0.126	0.332
	$\mathscr{E}(d)_{wri}$	0.125	0.123	0.127	0.128	0.151	0.122	0.134	0.136	0.131
Ours	nW	21	28	27	23	27	22	22	19	23.6
	$t(s)$	4.996	2.573	2.570	2.436	7.282	7.306	2.593	2.491	4.031

Table 2.1 – **Quantitative evaluation and comparison on synthetic images.** The table summarizes the 3D error ϵ_{3D} with the energy value $\mathscr{E}(d)$ in Eq.(2.10) for the baseline [51] and for our approach. Both metrics are measured on the full set of points denoted by *tot*, and just on the adjusted vertices, denoted by *wri*. We also show the number of detected wrinkles nW, and the computation budget $t(s)$ in seconds for our approach.

Finally, we introduce the time budget on a standard desktop computer Intel(R) Xenon(R) CPU ES-1620 v3 at 3.506GHz to resolve wrinkle detection, modeling, and estimation. When the CPU allows a parallel computation, the wrinkles are estimated in parallel, and its number does not drastically affect the total computation budget. On balance, it is worth mentioning that all the process just takes a few seconds on the commodity of a personal computer (see the last row in Tables2.1 and 2.2).

The differences between both estimations can be better perceived in a qualitative manner. Figure2.4 presents a qualitative evaluation and comparison of a few of the evaluated facial expressions between our method and the baseline [51]. As can be seen, our approach can recover the cheek lines strongly and other details around the facial geometry, which is not sufficiently recovered by current methods. Especially, it can be seen how the CNN-based solution fails to retrieve medium and fine details in some expressions (see first and third rows in Fig.2.4), even it produces large artifacts by setting points at one of the image corners (such as neutral, surprise and kiss expressions in the same figure). All these artifacts can be solved in the first stages of our approach in an efficient and effective manner.

Input		Img1	Img2	Img3	Img4	Img5	Img6
[51]	$\mathscr{E}(d)_{tot}$	0.2093	0.2543	2.2219	0.2969	0.3298	0.3067
	$\mathscr{E}(d)_{wri}$	0.1615	0.1810	0.2344	0.2109	0.3089	0.2130
Ours	$\mathscr{E}(d)_{tot}$	**0.2070**	**0.2505**	**2.2103**	**0.2929**	**0.3290**	**0.2968**
	$\mathscr{E}(d)_{wri}$	**0.1570**	**0.1729**	**0.2285**	**0.2002**	**0.3037**	**0.2101**
Ours	nW	307	16	21	73	124	18
	$t(s)$	7.844	5.3001	5.1339	6.8080	7.6930	4.6443

Table 2.2 – **Quantitative evaluation of our method on real images.** The table reports the photometric energy error $\mathscr{E}(d)$ (see Eq.(2.10)) for the baseline [51] and for our method approaching the full set of points and uniquely over the affected vertices to properly visualize the influence of wrinkles. As in the previous analysis, we also show the number of detected wrinkles nW and the computation time $t(s)$ in seconds, respectively for our approach. Images are denoted as Img1, Img2, Img3, Img4, Img5, and Img6 correspond to the first to sixth column in Fig.2.5.

2.6.2 Real Images

We now show the good performance of our framework applied to real images extracted from the internet. The set of images includes faces with a wide range of geometries, including subjects with different gender, ages, ethnic group, and facial gesture to prove the generality of our approach.

The comparison is carried with the energy values $\mathscr{E}(d)$ since no ground truth is available for these images. The procedure we follow is equivalent to the synthetic images tests, and we keep the CNN-based approach [51] as our baseline. Results for the full set of vertices and modified vertices can be appreciated in Table2.2. As we can recognize, the connection between the two values establishes an indirect metric to estimate the number of detections. For instance, in Img3 not all the wrinkles were properly captured due to noise (produced by filters added to the image for aesthetic purposes). The difference between values is high in comparison with other images. We can recognize how our method outperforms [51] toward all the metrics. Some examples from different viewpoints are represented in Fig.2.5. Again, it is worth pointing out that our method can obtain more accurate solutions than state of the art in terms of geometric details, as it can be observed in the different figures. Our method can recover different types of wrinkles, as well as scars or blood marks under different statuses (uncontrolled illumination, different resolution, and noise).

For completeness of our tests, we employ four indoor images for two subjects taken from [39]. As in the previous real data tests, we can only present a qualitative

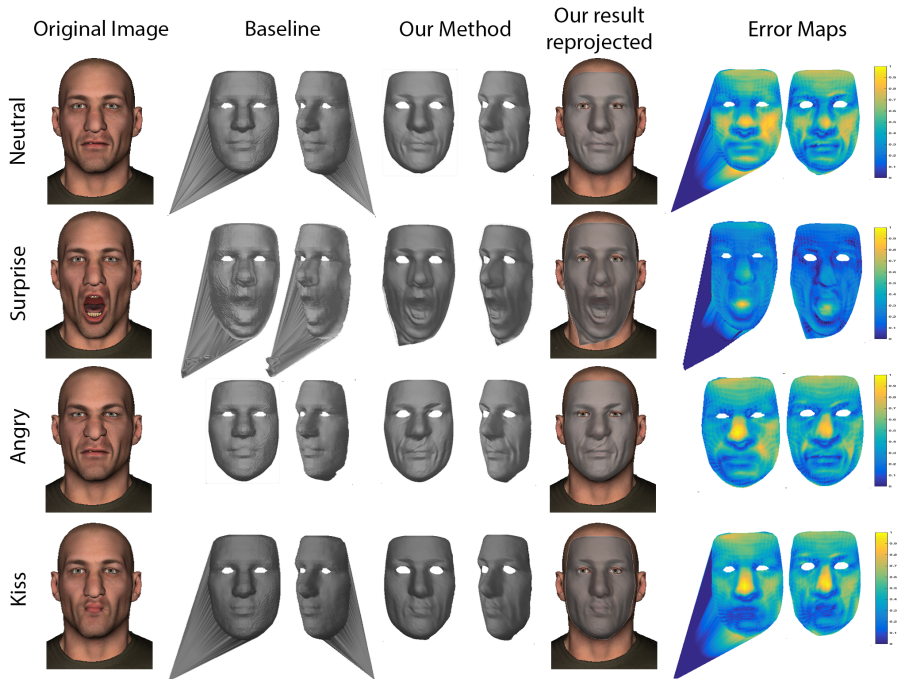


Figure 2.4 – **Detailed 3D face reconstruction from a single image on synthetic data.** We represent a qualitative evaluation and comparison of four different facial expressions, one per row. **First column:** Rendered synthetic image we use as input. **Second and Third columns:** Frontal and side views of the 3D face reconstruction we obtain by using [51]. Particularly, we use this estimation as an initialization. **Fourth and Fifth columns:** We display the same estimations after applying our formulation, which is also reprojected over the original image in the **Sixth column**. **Seventh and Eighth columns:** A vertex error map is represented between the baseline [51] and our estimation with respect to the 3D ground truth, respectively. As can be seen, our approach can recover a larger amount of fine details in 3D. Best viewed in color.

evaluation and comparison. In this test, we include the results provided by [39] which are available for this dataset. We display the analogous 3D reconstructions on Fig2.6. Despite other techniques exploiting temporal priors and a higher resolution in [39], our approach still appears to provide further detailed solutions than the rest of the evaluated methods, even when managing significantly fewer points.

In general terms, real images include more sophisticated facial features com-

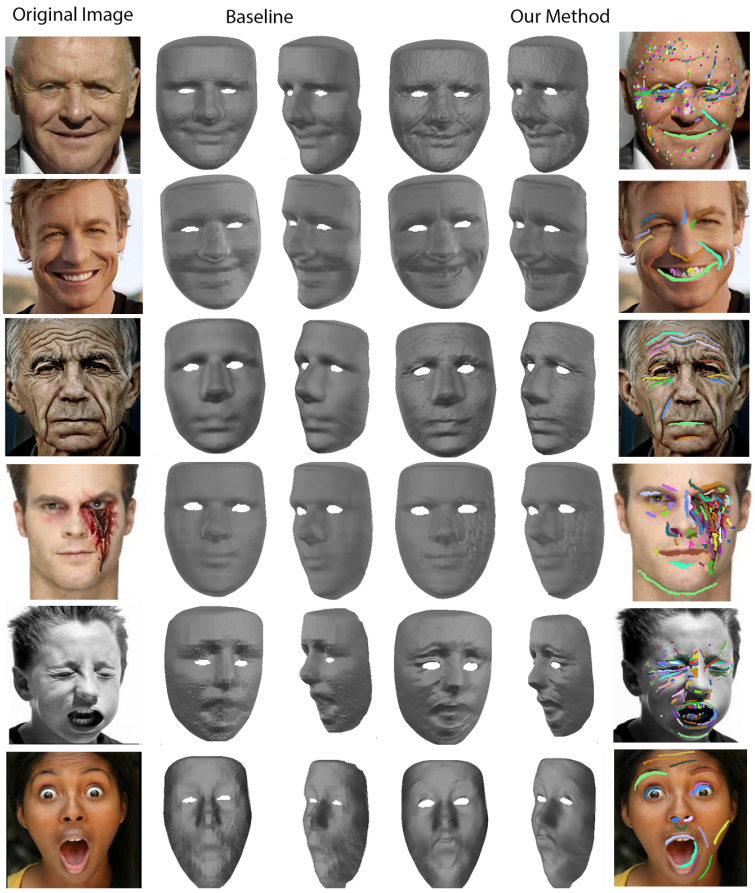


Figure 2.5 – **Detailed 3D face reconstruction from a single real image test**. Six different scenarios –varying age, gender, ethnic group, and facial expression– are displayed in rows. **First column:** Input image. **Second and third column:** Camera and side views of the 3D reconstruction obtained by [51]. **Fourth and fifth column:** Camera and side views of our 3D reconstruction. **Sixth column:** Detected wrinkles. Best viewed in color.

pared to synthetic data. It has to be considered that the design of realistic wrinkles still nowadays a challenging problem in 3D modeling and animation. Hence, the acquisition of realistic models from vision is significant to geometrically analyze

Figure 2.6 – **Detailed 3D reconstruction on real images**. Left and right display the same. **First row:** Four input images with different expressions. **From second to fourth row:** Reprojected mesh of the 3D reconstruction using [39], [51] and ours, respectively. Note that the solution provided by [39] includes a 200k-point, rather than using 20k points like us.

the local details. In this context, the input image resolution is a fundamental factor to obtain good results. As it is represented in Fig.2.7, we recognize a more fine acquisition when the image resolution increases.

2.7 Failure Cases

Finally, we would like to consider failure cases. Since our formulation relies on image partial derivatives to detect wrinkles, some conditions such as shadows or texture-varying areas can produce ambiguous situations. For instance, dark tattoos, strong cast shadows, and significant occlusions (i.e., dense beards) are some examples in which recovering detail becomes a laborious task, and our method may fail (observe some examples in Fig.2.8).

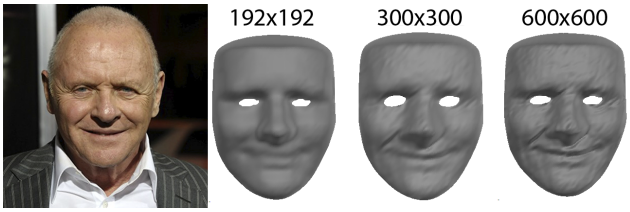


Figure 2.7 – **3D reconstruction as a function of the image resolution.** To reveal the impact of the image resolution above the final result, we run our approach on different down-sampled images. While most of the details are not recognized in low-resolution images, they become more precisely detected as the resolution in pictures increases.

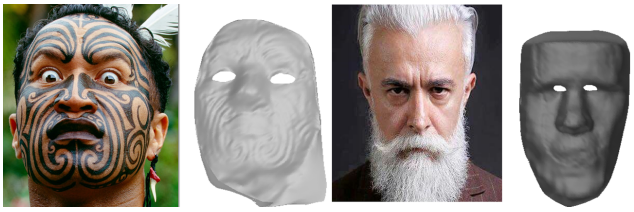


Figure 2.8 – **Failure cases**. Two ambiguous examples that our approach cannot resolve properly. **Left:** A dark tattooed Maori where our algorithm fails in distinguishing texture and shadow areas. **Right:** A thick beard produces a self-occlusion in a large region of the face. Although the estimation is visually correct, it does not consider nor recover the human beard.

2.8 Conclusion

In this chapter, we have proposed an intuitive and effective approach to retrieve detailed 3D reconstruction of faces from a single RGB image. Considering only the input image, our approach can obtain several features to parametrize the wrinkles without having been observed previously. This scheme allows us to model even person-specific attributes, such as scars or several shapes as a consequence of aging. Additionally, our approach is efficient since only need few seconds to solve the problem, by sorting out a photometric optimization problem. We have extensively evaluated our approach on both synthetic and real images, considering a wide range of variability in which we have outperformed existing state-of-the-art solutions. An interesting avenue for future research is to extend our formulation to handle more severe occlusions as well as a validation in real time at frame-rate.

3 Facial hair recovery from single RGB images

In this chapter, we present a novel energy-based framework that approaches the challenging problem of 3D reconstruction of facial hair from a single RGB image. We first identify hair pixels over the image through texture analysis and determine individual hair fibers. Next, we model the detected hair fibers in 3D using a parametric hair model based on 3D helixes. We propose to minimize energy formed by several terms, to adjust the hair parameters that better fit the image detections in 2D. The resulting hairs respond to the recovered fibers after a post-processing step where we encourage further realism. The presented approach generates realistic facial hair fibers from a single RGB image without requiring any training data nor user interaction. We provide an experimental evaluation of real-world pictures where several facial hairstyles and image conditions are observed, showing consistent results and establishing a comparison concerning competing approaches.

3.1 Motivation

Over the past few years, digital characters have been provided with a dash of growing realism, heightening the standards of the film and videogame industry. Besides the improvements in the skin, cloth, and illumination, hair has played an essential role in the building of convincing digital characters fulfilled with an identity. Pessig*et al.* [63] proffered the eyebrows are the most consistent feature for facial recognition. Unfortunately, retrieving the facial hair geometry from images or videos is a challenging task due to its structural complexity and variance between different subjects along with the different face areas. Although several image-based methods can create high-quality reconstructions [13, 52, 59, 61, 62, 84], they require specialized and expensive hardware to achieve the worth. As well they often demand a studio or a controlled environment. Moreover, except [13], none of them demonstrated any accomplishment with the recovery of facial hair.

In the recent past, diverse methods handled the hair reconstruction from a

single RGB image. Though, the first approaches [24, 26] demanded several user interactions. Especially [24] requires depth information and user annotations for the hair strands. Equivalently, [26] expects the user to determine sparse strokes to ascertain the local direction uncertainty. The user annotation dependence dropped to new data-driven solutions [49], which reduced the required volume of user interaction but not completely omitted. In a different direction, remarkable results were achieved with the use of structured light patterns [27] and electro-luminescent wires [46]. Unfortunately, any accomplishment was claimed over facial hair. Correspondingly, deep-learning approaches contributed to the field as hair growing direction estimation [25] and hair-style parametrization [90], amongst other techniques. Per contra, these approaches tend to be extraordinarily data demanding, while the active methods require expensive hardware settings.

In this chapter, we introduce our novel optimization framework that uses the texture information in an RGB image to estimate the facial hair fibers geometry in 3D. We exploit the orientation texture analysis methods [61] to detect the fibers and their orientation. Afterward, we step the pixel detection of the individual hair fibers, disjoining connections at significant orientation variances. Similar to [13], we commit the detection of hair crossings by connecting fibers close in the space with a comparable overall orientation. Hereafter, we estimate the parametrization of our hair generation model by minimizing a set of four different energies that consider different 2D detection properties. To develop the computation efficiency, we arranged the detected hair fibers in different groups according to their 2D properties (these are position, length, and orientation), obtaining a different parametrization per group. Eyelashes and eyebrows require to be studied independently due to the evident differences regarding beard and eyelash fibers. Finally, we heightened the final hair reconstruction realism by adding hair density and small random noise.

Our central contribution in this chapter is a model to grow 3D hair fibers over a recovered 3D face shape derived from a high resolution single RGB image. As a result, our method does not demand any additional information, user interaction, or training data. We prove the capability of our method across an extensive spectrum of facial hairstyles and facial geometries. Our approach is broadly evaluated on real high-quality RGB images from the Internet, demonstrating the appropriateness of our framework to reconstruct plausible 3D facial hairs right form pictures.

3.2 Related Work

Facial hair is one of the most fundamental features when it comes to facial realism. It fulfills digital characters with humanism and verisimilitude. Consequently, several methods addressed the description and reconstruction of hair from different

perspectives [13, 24, 52, 61, 84].

As has become the norm, the multiview approach is the most straightforward scheme. Since several points of view overcome the uncertainty in the reconstruction process. Among the first models, Paris et al. [61] implemented an image-based approach built on the study of the scattering properties in image sequences. It provided the capture of the hair fiber geometry based on these particular attributes. The premise was a highlight that has transcended to the present day. Even their results are outdated, their contribution describing the orientation fields of image scatterings was the basis for multiple methods. Wei et al. [84] prevailed on the image-based multiple view techniques. Their proposal was based on local hair orientations and was able to retrieve satisfactory results erasing almost every user interaction from previous approaches. The most distinguished contribution was their capacity to work under uncontrolled illumination conditions. Jakob et al. [52] presented a novel method able to accurately recover hair fibers from macro photographs with shallow depth of field by swapping the plane of focus along the hair volume. They found a similarity between focus plane swapping to multiple observations. Beeler et al. [13] presented a multiple view approach consisting of a coupled hair and skin multi-view stereo method based on images acquired in a controlled studio. This method was able to reconstruct not only the facial hairs but the epi surface where the hairs are generated from. Yu et al. [88] presented a hybrid image-CAD system to model hair strands. It works with 2 to 3 viewpoints, but requires the user to draw a few strokes to generate base 3D splines, Then, they connect the 3D orientations fields with the hair volume for growing hair fibers. In Zhang et al. [89], they introduce a novel four-view image-based method to model hair. First, they estimate the rough 3D shape using a 3D database of hair models. Then they synthesize the hair texture on the surface of the shape in the growing direction of the hair volume.

Other methods opted for hairstyle parametric formulations in monocular sequences and single image reconstructions. It is well known that these methods are underconstrained and need a priori information or strong assumptions to generate proper results. In the particular case of low-rank hairstyle models, realism is limited to the quality of the parameterized model. Luo et al. [59] introduced a method based on local coherent fibers and a novel graph structure. It studies the direction and the connectivity of the local fibers in the global structure. They further synthesize additional fibers to add realism to the final reconstruction. Hu et al. [48] presented a data-driven hair capture framework based on a fitting-algorithm. It included a voting-based decision step capable of growing hair segments structurally plausible based on a large dataset. Hu et al. [49] present a database of hair models with resources obtained over several repositories. Given a reference photo, with a few user guidance, they automatically select the best matching samples and combine

them in a consistent hairstyle with the user drawn stokes. Chai et al. [24] introduced a novel method with minimal user interaction. It consists of a parametric model-fitting method to reconstruct the face from a single image. Then, they perform the estimation of the facial hair via a Shape-from-Shading-based optimization method, which benefits from the inferred light from the face as well as depth cues, occlusions, and silhouettes. It further gets the advantage of an albedo prior, which specifies the most typical hair colors and occlusion variations.

Deep-learning formulation permitted the initial solutions from a single RGB image with any user interaction [25, 90]. Chai et al. [25] introduced a novel hierarchical deep neural network with no user interaction. It is able to perform both hair segmentation and growth direction estimation. They trained the model over a large annotated hair database with 3D models. Zhou et al. [90] presented a different deep neural network that automatically generates full 3D hair geometry from a single image. It commences with the 2D orientation fields and generates hair strand features. They consider hair collisions into their loss function to accomplish more plausible hairstyles. They use a large set of 3D hairstyles to train the network. It results in a guided interpolation within hairstyles given an image.

Structured light (SL) technology has achieved detailed and personalized avatars with high-quality results in both facial and full boy geometries. Chen et al. [27] proposed a novel method to capture hair based on four white structured light scanners. Their strip-edge coding algorithm is able to reconstruct hair with a pattern of 18 strips with accuracies of 1mm. However, their setup is complex, and a resolution on 1mm when it comes to human hair is insufficient. An original approach was introduced in [46], where a set of low-cost electroluminescent wires were proposed to be twisted into braided hair stand to illuminate from the inside and referencing them. The bypassed the requirements of hair texture or data-driven prior knowledge with their active 3D curves captured from several images. Even their results are genuine and reliable 3D hair braided strands, they rely on the use of specific and uncommon controlled lighting. Moreover, they require several images to reconstruct trustworthy outcomes.

While, model fitting approaches do not handle fiber-by-fiber retrieving, recalling these methods ineffective for accurate facial hair reconstruction. Further, they are demanding in data terms, especially deep-learning based methods. On the other hand, multi-view and active-light based methods require expensive or specific hardware and controlled illumination, reducing its applicability in real-world scenarios. Besides, most of the previous works had not demonstrated its suitability to reconstruct facial hair fibers but head hair fibers or hairstyles. We find in [13] an interesting exception. We present a similar concept but using a single-view image under uncontrolled lighting. Our method can handle hair fibers estimation from a good quality RGB image without requiring training data or other setups.

Algorithm 2 Recovery and parametrization of facial hairs from a single RGB image.

Coarse initialization with [51] $\mathbf{I} \rightarrow \mathbf{V}$
Volume to mesh conversion $\mathbf{V} \rightarrow \mathbf{S}$
Texture Analysis $\mathbf{I} \rightarrow (\mathbf{M},\mathbf{O})$
Hair detection from texture $(\mathbf{M},\mathbf{O}) \rightarrow \mathbf{H}$
Hair tracing and endpoint labeling $\mathbf{H} \rightarrow \mathbf{P}$
Hair clustering $\mathbf{P} \rightarrow \mathbf{P_k}$
Energy minimization $\boldsymbol{F}_i^h(\boldsymbol{p}^h, \boldsymbol{n}^h, l, w, r, \theta, g)$
Appending further realism $\boldsymbol{F}_i^h(\boldsymbol{p}^h, \boldsymbol{n}^h + \lambda_o, l + \lambda_l, w, r, \theta, g)$

3.3 Problem Formulation

In this section, we describe the computation of the 3D hair strands from a single RGB image. Including the early detection stages and further post-processing techniques to achieve further realism.

In a nutshell, our method first detects hair fibers in 2D via Gabor texture analysis. Then, based on the obtained orientations, we analyze and outline the full hair from its root to the tip at pixel-level. We group the detected hairs according to their 2D properties and location in clusters to take the maximum advantage of our optimization process. It is composed of four different parts, which in combination with the estimated 3D facial model [51], allows us to find the values of our parametric model per each of the groups. As a final adjustment, we supplement all the computed hairs with further realism by adding small random variations in their length and root orientation.

Figure3.1 illustrates a schematic of the overall approach, the connection amongst every part. In the remainder of this chapter, each step is illustrated in detail. Algorithm 2 offers a delineative visualization of the different steps.

3.4 Hair Detection in 2D

3.4.1 Texture Analysis

In this section, we have focused on the detection and description of hair properties in 2D with sole information directly extracted from the image. Our method does not demand any a priori information. On analyzing the image data, our first observation is that human hair may change in a considerable range of tones and shapes, making the problem harder to approach. We found the RGB space to be unsuitable to

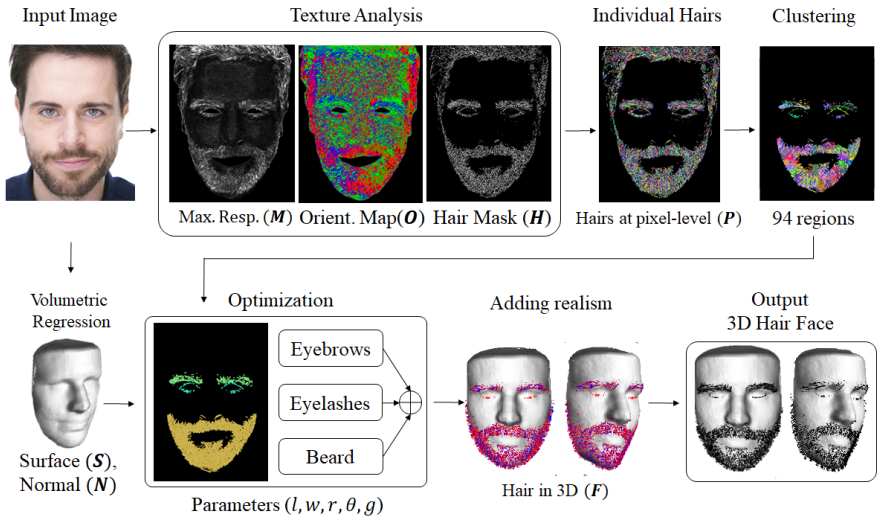


Figure 3.1 – **An overview of our pipeline**. From a single image, our approach first retrieves a 3D facial model (coded by $\boldsymbol{N}$ and $\boldsymbol{S}$) by applying a volumetric regression CNN approach [51]. Later, a hair map is detected over the image via Gabor texture analysis, where some attributes are obtained: the maximum response and the maximum orientation response (every orientation is represented by a different color) are denoted as $\boldsymbol{M}$ and $\boldsymbol{O}$, respectively, obtaining the final hair detection in the binary matrix $\boldsymbol{H}$. Next, we trace individual hair fibers via pixel-connectivity, orientation differences, and endpoints distance in $\boldsymbol{P}$. As different areas (beard and mustache, eyebrows, and eyelashes), need a different parametrization due to their variability, hair fibers are grouped in 94 different regions according to their location, orientation, and 2D length, which are included in one of the previous macro-classes. For these macro-classes, model parameters are estimated by optimization. Eventually, we append the results and add small random variations in the hair length and orientations as well as density to increase the realism. Red and blue lines represent the estimated and computed hairs, respectively.

correctly separate hair and skin since there are huge amounts of tones and shades of each. With this purpose, we consider the HSV color space as our best choice, since saturation and value channels provide us information which is independent of the skin and color hue. It allows us to exploit the greater uniformity of the hair fibers, together with a significant difference with regards to the skin pixels. The hue

channel is useful to identify the inner part of the eyes and the mouth where the texture is not analyzed. It is well known that these regions do not hold hair fibers. However, they can contain other features that can be confused with hair fibers in the texture analysis such as veins, or strong lip textures. For this, and similar to [13], we consider it convenient not to analyze the texture of these specific areas.

As stated in the work of [13, 61] hair can be successfully estimated in images via orientation analysis. Formally, they use a filter kernel K_θ for different θ orientations, at every 10 degrees, and keep the orientation that generates the largest score in the function:

$$\mathscr{F}(x,y) = |K_\theta * V|_{(x,y)} + |K_\theta * S|_{(x,y)}. \tag{3.1}$$

This function is applied at the pixel level so, at a pixel (x,y), for the value V and saturation S channels.

In a similar manner, we employ the real part of a Gabor filter bank consisting of 5 different wavelengths $\lambda = \{2, 2.5, 3, 3.5, 4\}$ and 18 orientations θ (from 0 to 170, at every 10 degrees). For each pixel on the image, we hold the maximum response of the filter $\boldsymbol{M}(x,y)$ and the orientation of the maximum response $\boldsymbol{O}(x,y)$ which later will allow us to detect individual hair fibers. Both are defined as:

$$\boldsymbol{M}(x,y) = \max(\mathscr{F}(x,y)), \tag{3.2}$$

$$\boldsymbol{O}(x,y) = \theta_{\max(\mathscr{F}(x,y))}. \tag{3.3}$$

This function is applied at the pixel level so, at a pixel (x,y), for the value V and saturation S channels.

Subsequently, we binarize the maximum response with a simple threshold of τ to reject low-confidence responses. With this filtered information, we define a binary hair mask $\boldsymbol{H}$ as:

$$\boldsymbol{H}(x,y) = \boldsymbol{M}(x,y) > \tau. \tag{3.4}$$

3.4.2 Individual hair trace

As we described in the introduction, the interest of this chapter is to effectively recover the geometry of facial hairs in 3D from a single RGB image. To this end, it is important to define a model to parametrize 3D hair fibers with parameters we can extract from the image jointly with the 2D properties. With this purpose, let us define a hair at a pixel level as $\boldsymbol{P}^h$, and the set of all the pixels in a hair as $\boldsymbol{p}_i^h \in \boldsymbol{P}^h$ with $\boldsymbol{p}_i^h = (x_i^h, y_i^h)$. The goal in this stage is to transform the different blobs in $\boldsymbol{H}$

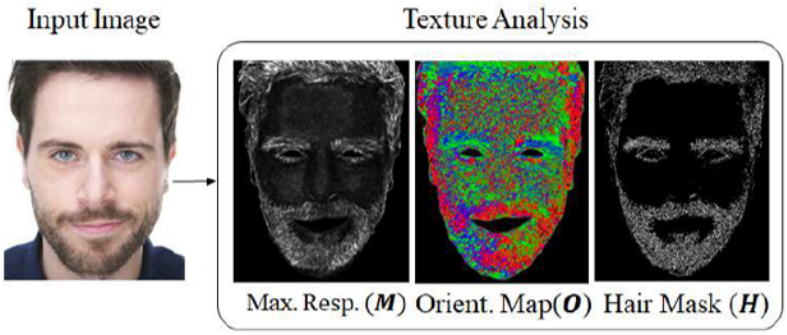


Figure 3.2 – **Hair detection process.** We detect hair via texture analysis. From an input image, we keep the maximum response of the filter bank **M** and the orientation of the maximum response **O**. With this data, we can filter the Maximum responses with a threshold and obtain the Hair Mask **H**.

into an ordered set of pixels $\boldsymbol{p}_i^h \in \boldsymbol{P}^h$ where $\boldsymbol{p}_0^h$ is the hair root and $\boldsymbol{p}_L^h$ the hair tip, where L denotes the length of the fiber.

With the previous definitions, we seek for hairs as 8-connected regions through the hair map $\boldsymbol{H}$. The hair constitution, crossings, and string shading strongly affect the solution since it may cause different detections grouped as a single detection and vice-versa. To determine a single hair fiber trace, we consider the orientation map $\boldsymbol{O}$ and ensure all the pixels in the same connected region has an orientation difference with the following pixel smaller than 10 degrees. We detach all the connections that do not accomplish this condition.

$$\boldsymbol{B} = |\boldsymbol{O}(\boldsymbol{p}_i^h) - \boldsymbol{O}(\boldsymbol{p}_{i+1}^h)| > 10, \tag{3.5}$$

where $\boldsymbol{B}$ represents the detached connections.

Accordingly, it would be beneficial to re-connect hair traces that were broken due to occlusions or self-shading. In [13], they pointed that hair fibers can be re-joined in 3D if they satisfy three conditions: 1) their endpoints are unconnected, 2) they are close in the space (or overlapped), and 3) the orientation variation is lower than 20 degrees. We established a similar setup in 2D, limiting the angle to 10 degrees instead of 20 degrees to be consistent with our previous step. We define the maximum distance limit for re-connect the endpoints equal to three pixels, which stands to be the average width of a hair fiber in an image. For all combinations of hairs, we re-join those fulfilling the previous conditions. We additionally allow more than two hairs to join in the same section if the overall of the segments satisfies the

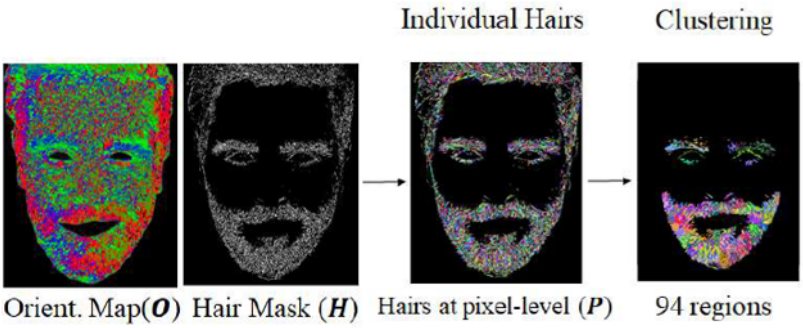


Figure 3.3 – **Individual hair tracing and clustering**. With the orientation of the maximum response **O** and the hair map **H**, we can approximate an individual hair trace, taking into account crossings and self-occlusions. For computational efficiency, we group the different hairs in 94 clusters to estimate the group parameters.

restrictions of both orientation and their endpoints are allowed to connect in a two-by-two association. The relevance of this step is imperative since the estimation of the resulting 3D hair properties is directly related to the measures extracted from visible 2D hair segments.

3.4.3 Endpoint labeling

In this section, we describe the process to define the hair direction labeling its endpoints as root and tip. This step is fundamental since our hair 3D model presented in the remaining of the chapter is implemented over the root position and grows in a specific manner to the tip location. To this end, we compute the facial landmarks with the method of [8]. These landmarks include the nose tip for beard, mustache, and eyebrows, and the averages of the eye landmarks as the central eye-points. Then, we define the hair growth direction labeling as the root the endpoint with the minimum Euclidean distance to the corresponding landmark:

$$\begin{aligned} \boldsymbol{p}^h_{root} &= \min(d(\boldsymbol{p}^h_i, (x^l_j, y^l_j))), \\ \boldsymbol{p}^h_{tip} &= \max(d(\boldsymbol{p}^h_i, (x^l_j, y^l_j))), \end{aligned} \tag{3.6}$$

where (x^l_j, y^l_j) denotes the landmarks. We employ the nose tip landmark to label the beard and mustache, the nose top for the eyebrows, and the average of the eye landmarks for the eyelashes. In case the endpoints were initially reversed, we shift the entire $\boldsymbol{P}^h$ values to satisfy the new endpoint labeling.

3.5 Hair modelling

In this stage, a new 3D hair is generated with the information extracted from the image. We opted to simulate the observations of real facial hair fibers as follows. When an incipient hair fiber grows, it accompanies the normal surface direction. When this fiber reaches a certain length it becomes ticker, and its trajectory is affected by other factors, for instance, the shaft weight, the gravity effect, and the follicle cross-section, which defines the fiber thickness and the curliness. It also affects the curvature in a two-dimensional plane, since the more burden the tip of the hair shaft supports, so the greatest is the gravity effect over the fiber. However, the curliness cannot be represented as deformation in a two-dimensional plane and requires to be extended to a 3D space as a local helix.

To consider all the mentioned variations, we present a hair model that ensures local coherence and smoothness. We define a hair fibers a 3D local helix representing both curliness and a gravity-like effect. Our model has five different parameters that establish the hair growing conditions and provide hair fiber-like results. These parameters determine the size (length l, and width w), the curliness (radius r, angle θ), and a gravity-like effect (g) that avoids shafts to lie suspended in the air. The last parameter, s, determines the resolution of the generated hairs. It is pre-defined by the user according to the requirements of the desired solution. The greater the value, the larger the resolution, and the slower to perform all the computations and visualizations. In our experiments, we chose a value of $s = 25$ for all the estimations. Figure 3.4 depicts the influence of the parameters in the enabling of the resulting hair fiber.

We consider $\boldsymbol{F}^h(\boldsymbol{p}^h, \boldsymbol{n}^h, l, w, r, \theta, g)$ the parametrization of a hair fiber, where $\boldsymbol{p}^h$ denotes the 3D position and $\boldsymbol{n}^h$ the 3D growing direction or hair orientation. Since the image is aligned with the initial 3D face mesh, the detections over the image are likewise aligned. With this advantage, for each root $\boldsymbol{p}_0^h$ we interpolate the three closets mesh vertices and obtain $\boldsymbol{p}^h$ as the mean of the positions and $\boldsymbol{n}^h$ as the average of the vertex normals. With the previous considerations and the parameters explained above we can define our facial hair model as:

$$\boldsymbol{F}_i^h(\boldsymbol{p}^h, \boldsymbol{n}^h, l, w, r, \theta, g) = \boldsymbol{p}_0^h + \boldsymbol{R}(\boldsymbol{n}^h)\frac{(i-1)\cdot l}{s} Hx(r, \theta, i) - Gy(g, i)\,, \tag{3.7}$$

where i denotes the i-th point in the hair fiber, $\boldsymbol{R}(\boldsymbol{n}^h)$ represents the rotation matrix, which adjusts the initial 3D helix's direction with the corresponding normal vector at the established position.

The 3D helix is determined over the x-axis as follows:

$$Hx(r, \theta, i) = (i-1, r\cdot\sin(\theta\cdot(i-1)), r\cdot\cos(\theta\cdot(i-1)))\,, \tag{3.8}$$

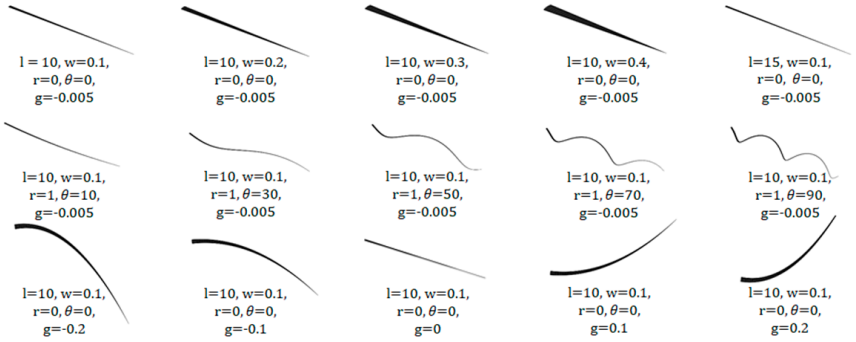


Figure 3.4 – **Parametric hair model**. Our parametric model depends on five parameters: length l, width w, the curliness parameters r and θ, and the gravity coefficient g. As it can be seen in the figure, thanks to our model we can obtain a wide variety of fibers. **Top:** Some instances varying l and w. **Middle:** Some instances as a function of the curliness parameters. **Bottom:** Modifying the gravity-like parameter.

where $Hx(r,\theta,i)$ designates the function providing the 3D coordinates of the 3D helix along the x axis at the i-th point. r is the radius and θ the angle between two consecutive points.

The gravity-like effect is estimated as follows:

$$Gy(g,i) = \boldsymbol{v} \cdot (i-1) \cdot (\cos(\alpha), -\frac{g(i-1)^2}{2}, \sin(\alpha)), \tag{3.9}$$

where $Gy(g,i)$ denotes the gravity-like effect at the i-th point of the helix. After several tests, we define $\boldsymbol{v} = [\frac{1}{\alpha}, 1, \frac{1}{\alpha}]$, being $\alpha = 10$ degrees as are the values that adapt the best the hair fibers behavior.

Finally, we compose a cylinder with two different radii over each segment of $\boldsymbol{F}^h$. Where the initial radius in $\boldsymbol{F}^h_0$ is defined beside the parameter w, and it decreases consistently unto the last point $\boldsymbol{F}^h_l$ where the cylinder radius is $w = 0$. The radii of the cylinder are directly related to the fiber width.

3.6 Energy Optimization

In this work, we take advantage of the local consistency of hair geometry, yet we consider it has different properties depending on face regions (beard/mustache, eyebrows, and eyelashes). For instance, eyebrows are not similar to beard fibers,

but each is similar within the same class, while thick and rumpled beards may have different properties even in the same class. To acknowledge these different behaves, we have established three separate optimization groups. The first applies to beard and mustache and clusters a total of 70 groups with similar length, position, and orientation via K-means. The second group covers the hairs belonging to the eyelashes, in total four clusters corresponding to each of the eye lines upper and lower of both eyes. The third and last group relates to the eyebrows, and it has ten clusters per eyebrow, likewise arranged with K-means. We optimize each of the 94 groups individually to satisfy the following energy minimization:

$$\mathscr{E}_{total} = \mathscr{E}_{len} + \mathscr{E}_{ori} + \mathscr{E}_{tip} + E_{cur}\,. \tag{3.10}$$

3.6.1 Length Term $\mathscr{E}_{len}$

The length energy term is a direct segment length comparison. We use the sum of Euclidean distances from the root pixel position $\boldsymbol{p}_0^h$ to the tip pixel position $\boldsymbol{p}_l^h$ and compare it against the corresponding distance in the xy-plane for the estimated hair root $\boldsymbol{f}_0^h$ and tip $\boldsymbol{f}_l^h$.

$$\mathscr{E}_{len} = \sum_h \|(\boldsymbol{p}_l^j - \boldsymbol{p}_0^h) - (\boldsymbol{f}_l^h - \boldsymbol{f}_0^h)\|_2^2\,. \tag{3.11}$$

3.6.2 Orientation Term $\mathscr{E}_{ori}$

It strengthens hair fibers to have comparable orientations than the disclosed in the given image. In practice, the global 3D orientation is defined by the surface normal vector, yet, the gravity-like parameter can force the fiber to grow in a different orientation.

$$\mathscr{E}_{ori} = \sum_h \|\tan^{-1}\left(\frac{p_{ly}^h - p_{0y}^h}{p_{lx}^h - p_{0x}^y}\right) - \tan^{-1}\left(\frac{f_{ly}^h - f_{0y}^h}{f_{lx}^h - f_{0x}^h}\right)\|_2^2\,. \tag{3.12}$$

3.6.3 Tip-to-tip Term $\mathscr{E}_{tip}$

Tip-to-tip cost encourages hair fibers to have the tip projection on the 2D plane approaching the tip detection in the image.

$$\mathscr{E}_{tip} = \sum_h \|(\boldsymbol{p}_l^h - \boldsymbol{f}_l^h)\|_2^2\,. \tag{3.13}$$

3.6.4 Curviness Term $\mathscr{E}_{cur}$

It limits the hair to match hardly with root and tip but miss in the remaining pixels. It computes the perpendicular distance from each fiber point to the closest point in the root-tip segment and examines the corresponding procedure with the detected hair on the image.

$$\mathscr{E}_{cur} = \sum_h \sum_i \| \frac{|(\boldsymbol{p}_l^h - \boldsymbol{p}_0^h) - (\boldsymbol{p}_0^h - \boldsymbol{p}_i^h)|}{(\boldsymbol{p}_l^h - \boldsymbol{p}_0^h)} - \frac{|(\boldsymbol{f}_l^h - \boldsymbol{f}_0^h) - (\boldsymbol{f}_0^h - \boldsymbol{f}_i^h)|}{(\boldsymbol{f}_l^h - \boldsymbol{f}_0^h)} \|_2^2. \tag{3.14}$$

3.6.5 Optimization

We solve Eq.(3.10) over the different groups in parallel, with the non-linear least-squares method. We apply a slight modification in the optimization of eyelashes and eyebrows. In the eyelash optimization, the procedure is equivalent though we force the detected roots to move along the spline formed by the eyelid landmarks. While in the eyebrows optimization, we force the hair fibers to grow in the tangent direction instead of the normal direction.

3.7 Further Realism

The next step consists in apply further realism to our recovered hair fibers since we found that estimating the hair parameters with large clusters, even when they have related properties, leads to losing significant realism. To overcome this issue, we present a two-step post-processing method consisting of adding density to the retrieved hair fibers and adding further verisimilitude by adding small random variations to the final 3D hairs.

3.7.1 Adding density

Exploring the individual fibers in 2D may lead to hair discard due to severe occlusion, incomplete tracking, or rejection of sections with poot connections. It heads to a density loss in the final result, which looks unrealistic. We propose a post-processing step that allows estimating the missing detections as well as adding further fibers to assemble the density of the image. To this end, we compare the binary mask generated by the orthographic projection of the computed hair elements $\pi(\boldsymbol{F})$, with the initial hair map $\boldsymbol{H}$. The greater the similarity between these elements, the greater the resemblance. To avoid false positives in the estimation, we expand the mask generated by $\pi(\boldsymbol{F})$ with morphological operators. Then we generate new hairs in the hair areas without any generated hairs. The number of required fibers is

computed as:

$$G = \sum_i \sum_j (\boldsymbol{H}(i,j) - \pi(\boldsymbol{F})(i,j)) \cdot (\pi(\boldsymbol{F}) \oplus \boldsymbol{D})(i,j)\,, \tag{3.15}$$

where G represents the number of possible pixels to grow a new hair, $\boldsymbol{D}$ is a binary 5×5 dilation mask, and $\oplus$ represents the binary dilation operator. G is a positive integer if the further density is required, and a negative integer if we added more than necessary hair fibers. If G is equal to zero, it implies that there is the exact amount of pixels in the hair map than projections of $\boldsymbol{F}$.

For each new required hair, we develop a new hair fiber $\boldsymbol{F}^k$ as a combination of the three closest hairs, taking into account the closest roots. To this end we average the parameters of the hair strand or the equivalent, averaging all the j-th points in $\boldsymbol{F}^k_j = \frac{1}{3}\sum_{i=1}^{3} \boldsymbol{F}^i_j$ where i denotes each of the three nearest neighbors. The process is iterated until G is equal or slightly lower to zero.

3.7.2 Adding small random variations

On the other hand, the estimation of an assortment of hairs with comparable parametrization leads to homogeneous hair along the face and consequently lack of realism. To overcome this problem, we combine small random noise to the resultant length parameter $\lambda_l \in [-0.05, 0.05]$ and a small random rotation amongst all the axes $\boldsymbol{\lambda}_r \in [-1, 1]$, where $\dim(\boldsymbol{\lambda}_\mathrm{r}) = 3$. We apply both variations $\boldsymbol{\lambda}_r$ and λ_l to the fiber estimated parameters to produce a hair shaft such as:

$$\boldsymbol{F}^h(\boldsymbol{p}^h, \boldsymbol{n}^h + \boldsymbol{\lambda}_r, l + \lambda_l, w, r, \theta, g)\,, \tag{3.16}$$

where $\boldsymbol{\lambda}_r$ adjusts the orientation given by $\boldsymbol{n}^h$ and λ_l the fiber length given by l.

3.8 Experimental Results

To properly validate our approach we require several types of images, including different hairstyles for both genders. We first obtained a significant bundle from Pexels platform. Additionally, we also present a qualitative comparison with respect to [13] using their data.

In our first experiment, we had first quantitatively evaluated on method with synthetic samples. We generated a total of 150 synthetic hairs over a half-sphere with our parametric hair model. Later, with the 2D coordinates of each point, we recover the 3D parameters employed with our optimization step. Finally, we compute the 3D error between points in the initial and the reconstructed hairs. To be consistent, we have used the configurations that were described in Fig.3.4. To

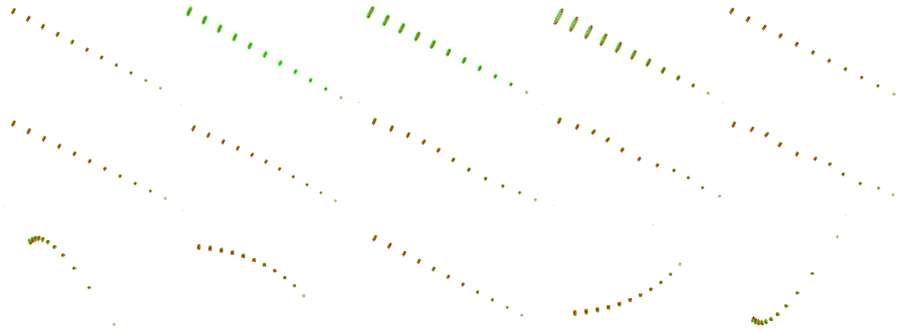


Figure 3.5 – **Qualitative evaluation of several hair fibers.** We depict the average amongst all the hairs in a single hair. Green circles represent the ground truth, and red dots our estimation. Best viewed in color.

(row,col)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
error 3D	0.027	0.004	0.006	0.014	0.017
time (s)	40.443	40.936	50.613	51.977	40.996
(row,col)	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)
error 3D	52.167	52.873	62.767	51.463	51.853
time (s)	0.014	0.013	0.017	0.015	0.013
(row,col)	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)
error 3D	0.002	0.002	0.006	0.003	0.008
time (s)	40.443	59.267	61.289	61.311	64.152

Table 3.1 – **Quantitative evaluation of several hair fibers and time budget.** 3D errors of the hair fibers depicted in Fig.3.5, and the corresponding computation time in seconds. Each row block represents a row in the figure. The avewrage 3D error is 0.011 and the average computation time 52.170 s.

this end, we report in Table3.1 the 3D errors as an average of the 3D error amongst all the points. Qualitatively, Figure3.5 depicts the average hair, revealing where these errors are located. As can be observed, in both cases our method achieves competing results, providing accurate hair reconstructions.

Our second set of experiments tests the effectiveness of our method on real scenarios. In this case, we consider several images, from short beards to full beards and mustaches. We also consider eyebrows and eyelashes. A representative set of results is displayed on Fig.3.6, where we can observe how our method produces

Image	1	2	3	4	5	6
im res.	849×1273	1024×768	1200×825	1600×2400	4000×6000	393×588
hairs up	766(61)	62(2)	72(6)	181(19)	804(6)	168(13)
hairs lo	10251(510)	9575(155)	17874(408)	4647(41)	0(0)	0(0)

Table 3.2 – **Number of reconstructed hair fibers for pictures on Fig.3.6**. We report for every picture on Fig.3.6, its resolution, and the number of retrieved hairs on the upper/lower parts of the face, showing in parenthesis the added hairs in the post-processing step. To this end, we consider the location of every picture on the figure, indicating its row and column position.

Image	1	2	3	4	5	6
im res.	1265×1920	845×650	1760×2640	1750×1168	3999×3999	5075×5760
hairs up	551(63)	609(55)	66(9)	334(67)	7132(238)	162(7)
hairs lo	6250 (212)	7332(746)	7733(440)	9384(361)	0(0)	0(0)

Table 3.3 – **Number of reconstructed hair fibers for pictures on Fig.3.7**. We report for every picture on Fig.3.7, its resolution, and the number of retrieved hairs on the upper/lower parts of the face, showing in parenthesis the added hairs in the post-processing step. To this end, we consider the location of every picture on the figure, indicating its row and column position.

realistic solutions on challenging scenarios with short and long beards, mustache, eyebrows, and eyelashes. Moreover, it achieves satisfactory results with partial occlusions, as self-occluding beards (see subjects (2,1) and (3,1) on the previous figure). We also report some numbers regarding these experiments in Table3.2, where the number of hair fibers is included. It is worth noting our approach can recover a large number of hair fibers in different areas.

Considering subjects with eyebrows and eyelashes, we include a challenging case concerning a subject with eyeglasses. Expressly, the subject (5,2) represents a challenging scenario due to the poor texture produced by makeup. Opportunely, our approach can recoversmall pieces of eyebrow hair instead of full hairs owing to the large image resolution and quality. Different example, it is the subject (6,1), Frida Kahlo, where our approach is evaluated for a low-resolution picture. As in the previous case, our approach also produces a visually realistic solution. Numeric validation of these experiments are reported in Table3.2. In Fig.3.8 we include some detailed close-ups, where it can be observed the realism we achieve with our approach.

In all cases, we use un-optimized Matlab code on an Intel(R) Xenon(R) CPU

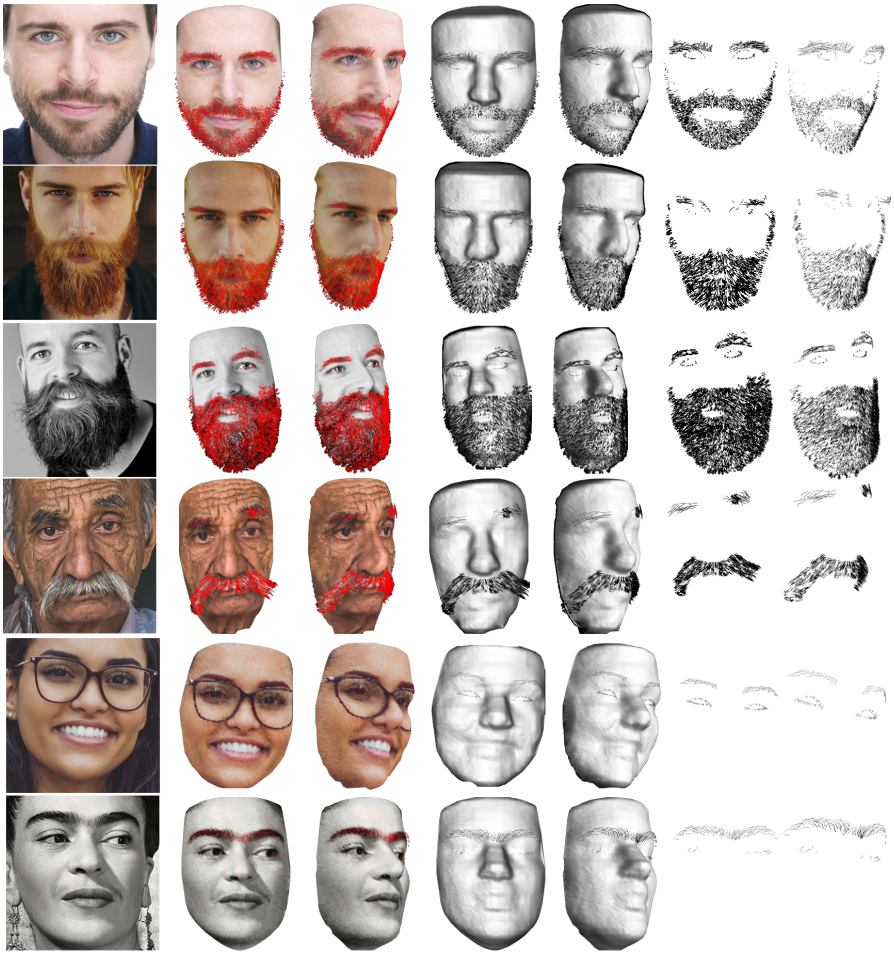


Figure 3.6 – **Face reconstruction with different facial hair styles**. **First column:** Input image. **Second and third columns:** Frontal and side views of our 3D hair+face reconstruction over a textured face. The hair fibers are represented by red lines. **Fourth and fifth columns:** Frontal and side views of our estimated geometry, without considering any texture. **Sixth and seventh columns:** Just observing our hair estimation.

ES-1620 v3 at 3.506GHz. The full pipeline run-time depends on the amount and

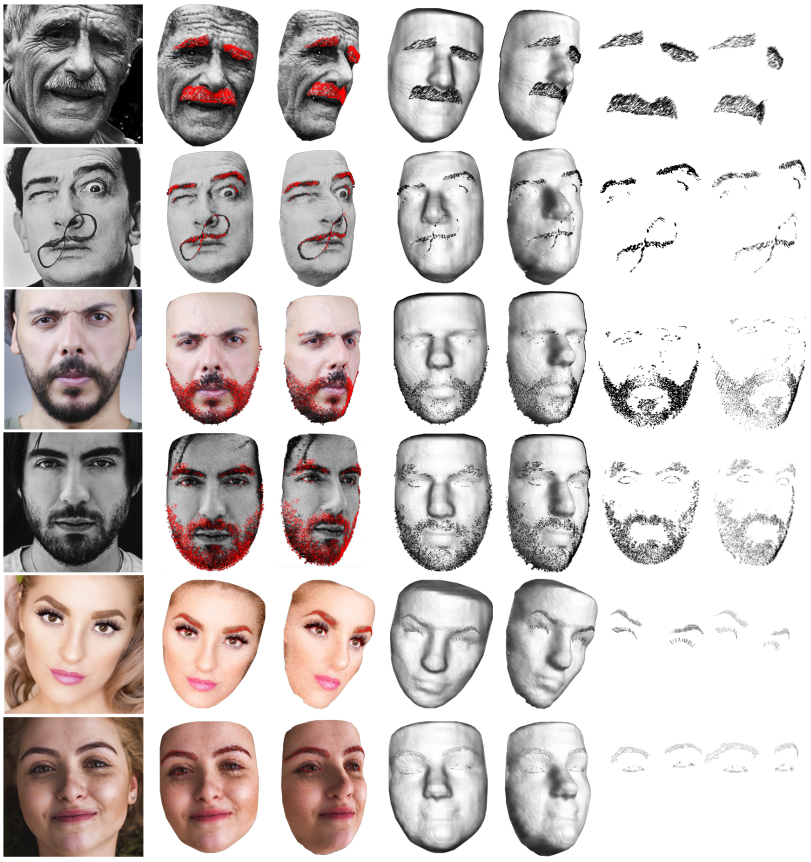


Figure 3.7 – **Face reconstruction with different facial hair styles**. We represent the same information than in figure 3.6 with different data. **First column:** Input image. **Second and third columns:** Frontal and side views of our 3D hair+face reconstruction over a textured face. The hair fibers are represented by red lines. **Fourth and fifth columns:** Frontal and side views of our estimated geometry, without considering any texture. **Sixth and seventh columns:** Just observing our hair estimation.

complexity of the hairs to recover. It takes from 15 minutes to recover examples with upper hair only to 10 hours to recover the subject (3,1) in Fig.3.6, where the hair density is extremely large.

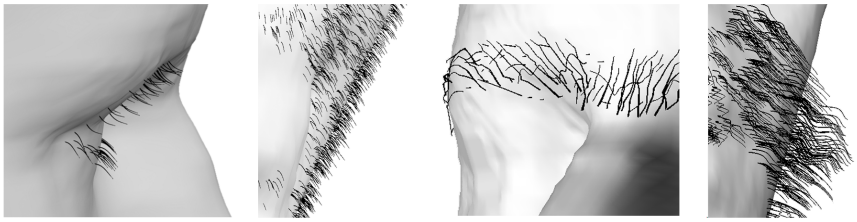


Figure 3.8 – **Close-up results**. Some close-ups of detailed instances are displayed. **First and second column:** eyelashes and a piece of beard around the mouth are represented for the subject (1,1) on Fig.3.6. **Third column:** unveils the thick eyebrow of Frida Kahlo, subject (6,1). **Fourth column:** represents a man's mustache, picture (4,1). In all cases, we can observe how the hair fibers are successfully recovered, and they are visually coherent.

As a last experiment, we compare our method with [13] in a qualitative manner. It is worth mentioning that our approach only requires an RGB image under general and uncontrolled lighting conditions as input, while [13] needs a calibrated multi-camera system. Despite this disadvantage in terms of hardware resources, our approach achieves competing results (see Fig.3.9 for a qualitative comparison). As it can be discerned, our method is effective in locating the hair fibers and optimizing their parametrization in the face area.

Our method offers an excelent performance working with a single image in both hair catpure and hair fiber recovery. However, our method has limitations and works better in better in the presence of short and scattered hairs on high-resolution pictures. We similarly find helpful to work with noiseless images on faces without make up or other texture disturbing elements, since the contrast between hair and skin is the most clear. Although our method can reconstruct like-wise facial hairs when the previous situations are not given. We also found the hair fibers with big orientation changes are difficult to be recovered (see Dali's example in Fig.3.6), since our approach is not able to fully trace individual hairs with angles larger than 10 degrees. Furthermore, the hair must be obvious at pixel-level to be recognized.

3.9 Conclusion

In this chapter, we have introduced our framework that geometrically recovers 3D facial hair fibers from a single RGB image without any training data. Accordingly, we have proposed a facial hair parametric model based on 3D helixes and a gravity-

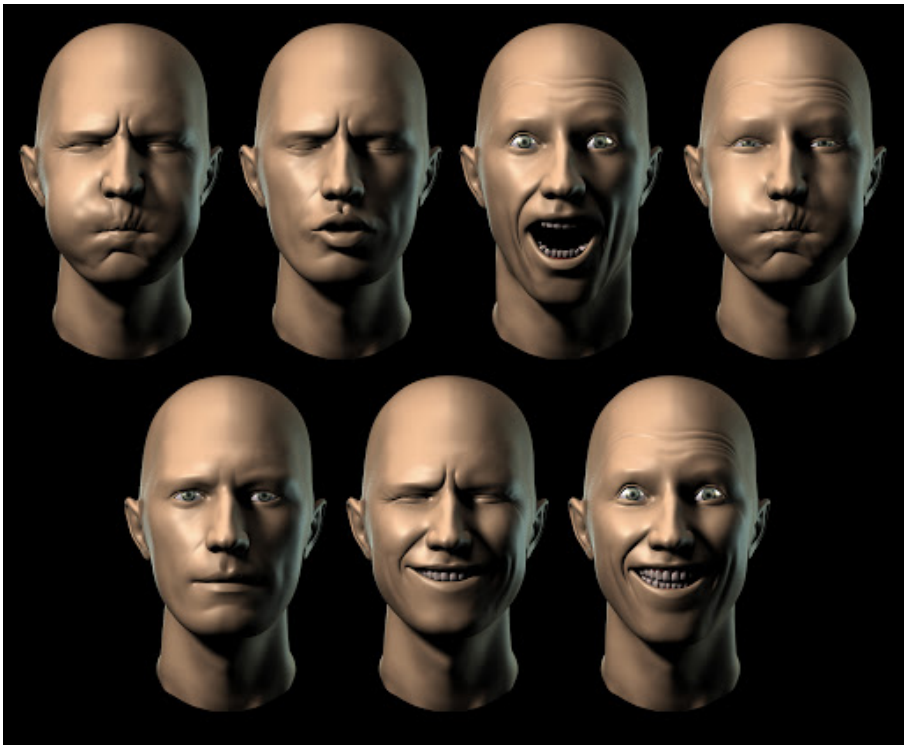
Figure 3.9 – **Qualitative comparison on 3D hair+face reconstruction. First column:** input RGB image for our approach. It is worth noting that the solution in [13] requires 14 cameras along with 4 flashes, i.e., a very constrained calibration is demanded. **Second and third column:** frontal and side views using [13]. **Fourth and fifth column:** our solution.

like effect to simulate the best the different facial hair configurations found on the images. To fit the previous model, we presented a set of energies that rely on 2D hair detections over different face areas to estimate the parameters directly from the image. Furthermore, our method does not demand training data, user interaction, or any particular setup.

We have validated our approach over a comprehensive collection of images with uncontrolled illumination and presented consistent and realistic results, including challenging cases as thick beards, eyeglasses, low-resolution pictures, and eyebrow makeup. Additionally, we contrast our approach with the current state of the art, where our method retrieves competing results despite the clear disadvantage in terms of hardware and single-image versus multiview studio setups.

We recognize the facial hair as the most critical part of a realistic reconstructed face. In some cases, like eyelashes, this detail is subtle but essential to get right. For this reason, future research lines are to join different procedures to retrieve various aspects of facial hair without delimiting the face area, but the full head structure including the neck, which is a plausible spot for men to have hair.

Expressions Part II



Blending shapes expressions developed by the digital artist Francesco Lupo [60].

4 2D-to-3D Facial Expression Transfer

Remodeling the expression and physical features of a face from an input image or video is a common topic in the 2D domain. During the last years, several methods have been proposed to solve this problem. However, in this paper, we bring this problem one step further and propose a 2D-to-3D framework. Given an input RGB video of a human face under neutral expression, it remodels the face to a potentially non-observed expression. For this purpose, we parameterize the rest shape –obtained from standard factorization approaches over the input video– using a triangular mesh which is further clustered into larger macro-segment with identical semantic representation as face areas. The expression transfer problem is then posed as a direct mapping between the shape containing the expression, for instance, a blend shapes of an off-the-shelf 3D dataset of human facial expressions, and the source shape. We resolve the mapping to be geometrically consistent between different 3D models, since we require points in a specific face region, to map on semantic equivalent regions. We validate our approach on several synthetic and real examples of input faces that widely differ from the source shapes, yielding very realistic expression transfers even in cases with topology changes, such as a synthetic video sequence of a single-eyed cyclops.

4.1 Motivation

The acquisition of facial expressions via time-varying 3D face models has grown considerably as a technology to transfer human face deformations to virtual avatars in movies and video games. First motion capture methods were based on optical markers, placed over the body and face. They allowed capturing a sparse representation of the human motion. The optical markers gave a very precise detection, however, besides being intrusive, these methods could not obtain accurate dense reconstructions, and as a consequence, small details as subtle gestures are not handled. Recently, dense face expression transfer has been addressed from a

deep-learning perspective, by learning 2D-to-2D mappings from large amounts of training images [9, 31, 65].

In this chapter, we introduce an approach that moves beyond the 2D-to-2D methods. Our solution densely transfers face expressions in the 3D domain. It exploits the potential of the so-called structure-from-motion algorithms and does not require training data at all nor placing markers on the actor's face. Particularly, given an input RGB video of a person under a neutral or challenging expression, we first leverage on a standard structure from motion framework to estimate its 3D face shape. On the side of the expression source, we assume it is provided for transferring via a low-rank 3D dataset of expressions composed by shape at rest and a determined number of blending shapes under specific expressions. We validate with both synthetic datasets and real-world facial expressions extracted from a video.

First, we develop a direct mapping system to match both faces at rest. In this mapping, we address several challenges, including different facial topologies, distinct mesh resolutions, and noise in the input shape, especially in the face from video sets. We overcome these difficulties by identifying common semantic regions defined by facial landmarks. These regions allow to locally solving the mapping for each of the regions, reducing the computation time and forcing the result to be semantically consistent. With the resolved mapping, we can transfer a wide spectrum of 3D expressions to the input face. See the overview of the presented approach in Fig.4.1.

We broadly evaluate and demonstrate the effectiveness of our method on a wide number of synthetic and real datasets, considering different mesh resolutions and geometries, even cases with large topology changes between the input and the source shapes. For instance, we show that the expressions of our low-rank 3D dataset can be transferred to a single-eyed face of a cyclops.

4.2 Related Work

Human motion capture, more specifically facial performance capture, has been extensively studied during the past years [64] [38] [70]. Originally, this problem was addressed with optical markers [14] or light patterns [85] to simplify the tracking process. However, these methods are intrusive and sparse. This restricts the acquisition of high-frequency deformations details like expression wrinkles and skin folds.

Otherwise, markerless approaches present a non-intrusive alternative, but tracking is not robust and may fail due to fast motion or complex deformations. Foremost amongst these methods are linear [20], or multilinear [19, 82] models, multiview-

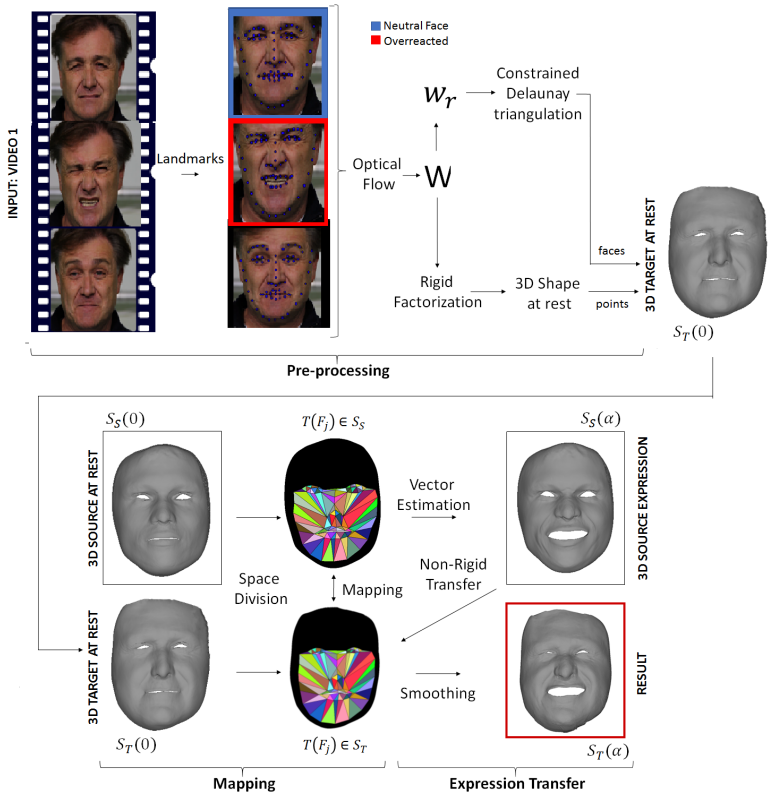


Figure 4.1 – **Overview of our expression-transfer approach**. Our approach consists of three stages: video to 3D shape with landmarks, mapping, and finally an expression transfer with smoothing. In the first stage (see the left part in the picture), a 3D shape from the optical flow is computed by rigid factorization, where overreacted measurements can be automatically removed to guarantee convergence, and a reference frame is selected. After that, we perform a sub-region mapping stage (represented in the middle), considering both resting target and the corresponding source face. In both cases, the shape is split into subregions to define a 3D-to-3D mapping between surfaces, and fitting the model. Finally, we perform the expression transfer stage (see right part) where the 3D configuration of a specific expression is transferred from the 3D source to the 3D target face model.

stereo solutions [12, 17], and techniques based on RGB-D data [16, 76]. Most previous approaches have focused on recovering a small set of parameters of a predefined low-rank model. Consequently, they are not suitable to retrieve detailed face deformations.

Outstanding results were achieved in dense 2D-to-2D facial expression transfer [9, 65, 76], and 3D-to-3D [72]. In this context, deep-learning approaches have been firstly introduced in this area [32] with prominent image-to-image results.

Recently, the 2D-to-3D approach [77] shows striking results, but due to the nature of their formulation, it is unable to retrieve fine-details, and its applicability is limited to the expressions lying in a linear shape subspace with known rank. In this chapter we propose a new mapping model that solves the previous limitations, by removing the need for a predefined low-rank model for the source reconstruction and the identification of a large amount of training data.

4.3 Preliminars

In this chapter, given a video of a face with potentially non-rigid expressions, we estimate the 3D reconstruction of the base face and a set of specified and non-observed expressions given by a template. In the first stage, we consider the specific expression is included in a template set and its selection is assumed to be known. In the second stage, we consider the on-line estimation of the template expression from a different face source, then the expression is captured in one of the faces and transferred to the acquired 3D target model.

We have focused on the development of our method of producing accurate and robust results. It is assumed that the template input is provided, and the sensor capturing the source face has quality enough to compute the pixel correspondences between the different frames of the input video via optical flow.

Let us consider an expression set of b expressions in 3D with their corresponding shape at rest. Every face in the set is made of p 3D points, represented by the matrix $\mathbf{S}_S(\alpha) = [\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_p]$, where the i-th column includes the 3D coordinates for the point $\mathbf{s}_i \in \mathbb{R}^3$. The parameter $\alpha = \{0, \ldots, b\}$ denotes a basis index, being $\alpha = 0$ for the shape at rest. Every basis denotes an expression basis.

We propose a different geometric representation based on triangular patches T, where each vertex in 3D corresponds to one of the triangular patches and can be represented as a set of barycentric coordinates and a displacement in the normal direction of the triangle. In a similar manner, we can define now a target face shape composed of n 3D points, represented by the matrix $\mathbf{S}_T(\alpha) = [\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n]$, which can be discretized into R triangular elements. In both cases, the parameter α can take a value from 0 to b, reserving the entry of 0 to indicate our input estimation or

shape at rest.

The challenging part of this work is to automatically estimate the 2D-to-3D mapping between the observed face in the video, and the target face with a specific expression. All without requiring any training data to constrain the solution, neither the specification of a pre-defined shape basis to project the solution. It is worth pointing out that our expression-transfer algorithm can even work when the resolution of both shapes $\mathbf{S}_S(\alpha)$ and $\mathbf{S}_T(\alpha)$ are different, i.e., when the number of points and triangular faces differs.

To this end, we introduce an efficient algorithm which works in three progressive stages: an extended structure-from-motion stage, to obtain a 3D estimation from 2D that considers both sparse and dense solutions; a step to establish the mapping between the target and the source shape; and finally, an expression-transfer stage that allows recovering the 3D estimation with a specific expression. Fig.4.1 depicts a summary of our approach. Next, we explain in deeper every stage of our method.

4.4 From RGB video to 3D Model

The initial stage consists of recovering the 3D shape at rest model from monocular video input. No further information about the video is required, and the observed face can potentially undergo non-rigid motions. In the last years, this problem has been addressed by non-rigid structure from motion approaches [2, 3, 4], showing accurate results even with incomplete 2D point tracks. However, our approach only requires a 3D model with a shape at rest, rather than knowing a 3D shape per image frame. For this reason, the resulting reconstruction can be easily computed by a rigid structure from motion techniques [79] with thep previous filtering of the non-rigid deformations.

We use a collection of 68 landmarks denoted as $\mathscr{F}$ extracted along the monocular video with OpenFace [8]. These observations are then analyzed to find the image frames where the neutral expressions appear, assigning the most neutral expression as the reference frame. This procedure is performed via Procrustes analysis of every frame against a 2D model extracted from our source $\mathbf{S}_S(0)$. The resulting values allow to determine which frames are neutral in terms of deformation or those that are overreacted and may make the process fail. We select those with an error lower than a threshold and discard those with a larger value, taking the frame as the one with the lowest error. Finding this frame is a key factor in our approach, since over this frame we compute a facial mask, and establish a reference to estimate optical flow.

After that, we apply state-of-the-art dense optical flow [37] to obtain 2D trajectories in the monocular video, collecting all the observations in the measurement

Algorithm 3 Algorithm including the main steps of facial esptression transfer.

Obtain facial landmarks $\mathscr{L}$ via [8]
Remove overreacted frames
Compute optical flow $\mathbf{W}$
Mesh aligning with landmarks $\mathbf{S}_S(0), \mathbf{S}_T(0)$
Low-resolution correspondences $\mathscr{T}(\mathscr{L})$ with Eq. (4.2)
KNN of unasigned vertices
Super-resolution correspondences $C_S(\mathscr{L})$ with Eq. (4.2)
Mapping of expression vectors $\mathscr{M}$ and $\mathscr{M}^{-1}$ with Eq. (4.5) and Eq. (4.6)
Transferring the expression vectors with Eq. (4.7) Smoothing of the resultant vector set with Eq. (4.8)

matrix $\mathbf{W} \in \mathbb{R}^{2f \times n}$, where f is the number of frames and n the number of points. As last step, we infer the 3D model from the measurement matrix via matrix decomposition. In general terms, we could apply recent works on non-rigid reconstruction and exploit the estimated time-varying shape to compute a mean shape. However, these algorithms can become computationally demanding for dense scenarios. To solve this limitation, we rely on rigid [79] approaches, since it is well-known the 3D estimation these methods produce is accurate enough when the motion is rigid dominant, as it is the case for face acquisition. We have observed experimentally the mean shape by applying rigid factorization after filtering overreacted expresions is roughly the same as that estimated with non-rigid approaches, showing robustness on the estimations.

For later computations, we convert the point cloud into triangulated meshes. To this end, we obtain the point connectivity over the 2D reference frame, where the connection is easily defined, applying a Delaunay triangulation [29]. For simplicity, we have used the Delaunay triangulation, although we could take advantage of having an estimation of the 3D rest shape and easily use alternative connectivity algorithms.

4.5 Mapping function

This section explains the second stage of our method where our goal is to establish a mapping function between the source and target models. Recall that this problem is not direct since the number of points and triangles to define every shape can be different.

To establish a correspondence with local coherence, we divide every face shape

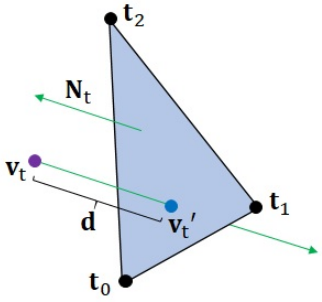


Figure 4.2 – **Graphical representation of the input-output funcion**.

through 101 triangular pieces (some pieces can be seen in Fig.4.3), grouping the points with a similar topology. These large triangles are defined by $T(\mathscr{L})$ where T denotes a super-triangular piece defined by three points of $\mathscr{L}$. With this representation, we can simplify the process of comparing faces of different topologies (races, genders, etc.) whose geometric information is also different, since we compare each of the super-triangular piece in the source face, with the corresponding piece of thetarget face .

Our model $\mathscr{M}$ is unsupervised and can classify each point on the 3D mesh to their corresponding subregion. We perform a corrective classification step which includes an iterative KNN to correct all the points that cannot be labeled in the first step. The iterative KNN labels the dissident points according to the values of their direct neighbors, but to prevent the shape change in the groups, we do not label a point if it has not at least K direct neighbors assigned. With this condition, we will iterate until enough reliable information is available to label the points. Lastly, we define the mapping model in which we estimate the point correspondence between the two geometries by generalized barycentric parametrization techniques.

4.5.1 Point subregion classification

The model we propose consists of a parametric mapping function defined by:

$$\mathscr{M} : \mathbf{S}_S(\mathscr{P}) \mapsto \mathbf{S}_T(\mathscr{P}), \tag{4.1}$$

where $\mathscr{P}$ represents a shape parametrization, and $\mathbf{S}_S(\mathscr{P})$ is known for the full set of $\mathscr{P}$, while $\mathbf{S}_T(\mathscr{P})$ is only known at the shape-at-rest parametrization $\mathbf{S}_T(0)$. We compute the 2D projection over the XY plane of the 3D points $\mathbf{S}_T(0)$ and the triangles $T(\mathscr{L})$. Then, we test for all the possible pieces $T(\mathscr{L})$ and all the points in $\mathbf{S}_S(0)$, if a

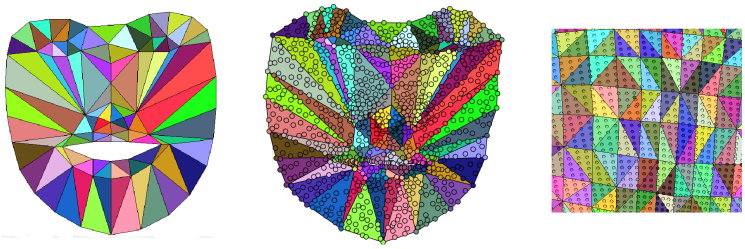


Figure 4.3 – **Graphical representation of the coarse and fine point-triangle classifications**. **Left:** The 101 triangular regions of a template face. **Center:** The different vertex points according to the 101 area coarse classifications on a source face. Each point is represented with the same color code as the belonging region. **Right:** we can observe a patch containing the different mesh fine triangles with the same point-triangle color codification.

point $v_t^i \in \mathbf{S}_T(0)$ lies in any triangle. To find out if v_t^i lies in the triangle $T(\mathscr{L}^j)$ we define the following inside-outside function:

$$\mu(v_t^i, t_j^n) = \begin{cases} 1, & if\ \mathbf{N}((t_j^{n+1} - t_j^n) \times (v_t^i - t_j^n)) \geq 0 \\ 0, & otherwise \end{cases}, \tag{4.2}$$

where t_j^0, t_j^1 and t_j^2 represent the vertices of the triangle $T(\mathscr{L}^j)$, and $\mathbf{N}$ is the corresponding normal vector.

The final decision is determined by the sum of the different values of $\mu(v_t^i, t_j^n)$ for every pair designed by the point v_t^i and triangle vertex. If the sum equals to 3, the point v_t^i lies inside the triangle $T(\mathscr{L}^j)$ and can be defined by barycentric coordinates Λ_T. Similarly, we determine the triangulas element of each of the triangular elements in the source face $\mu(v_s^i, t_j^n)$ and obtain Λ_S. In this case, an element can belong to more than one subregion, for example, if a triangle has each of its vertices on different subregions the triangle will belong to all these regions, so by its geometry, it can belong to three subregions.

Due to the precision of computation, we need to correct some points which remain unassigned. We resolve this ambiguity with an iterative clustering process using K-Nearest Neighbors (KNN), in which we assign to every unassigned point the most plausible value rely upon their KNN. It must be iterative due to the configuration of the clusters which are triangles instead of ellipses. So for each iteration of the KNN, we just assign a point if at least it has K labeled direct neighbors. If not, the point remains unassigned until enough reliable information is available, thus

the triangular shape of the clusters remains intact.

4.5.2 Model Fitting

After applying the point subregion clasification, we can achieve a finer definition over a smaller element of $\mathbf{S}_S(0)$. Let C_S be the set of all the elements of $\mathbf{S}_S(0)$ where c_S^k represents a single element of C_S. By the previous step, we can know that an element belongs to a subregion (or more), so $c_S^k \in T(\mathscr{L}_j)$. As we mentioned previously, we can represent the 3D point in barycentric coordinates, so now we have $\lambda_t^i \equiv v_t^i$. If we project the point v_t^i over each plane $C_S(\mathscr{L}_j)$ we minimize the distance of the point over the plane projection $d = \mathbf{N}(v_i - t_j^n)$, subject to the condition that the projection lies in the triangle ($v_i^s \in c_S^k$).

The previous parametrization of the point allows approximating it to the parametric solution, $\mathbf{T}^{\mathscr{L}_j}$ and $\mathbf{T}^{C_T^k}$ encode the edges of the triangles $T(\mathscr{L}^i)$ and C_T^k, respectively:

$$\mathbf{T}^{\mathscr{L}_j} = \begin{bmatrix} t_{x0}^j - t_{x2}^j & t_{x1}^j - t_{x2}^j \\ t_{y0}^j - t_{y2}^j & t_{y1}^j - t_{y2}^j \\ t_{z0}^j - t_{z2}^j & t_{z1}^j - t_{z2}^j \end{bmatrix}, \tag{4.3}$$

$$\mathbf{T}^{C_T^k} = \begin{bmatrix} c_{x0}^k - c_{x2}^k & c_{x1}^k - c_{x2}^k \\ c_{y0}^k - c_{y2}^k & c_{y1}^k - c_{y2}^k \\ c_{z0}^k - c_{z2}^k & c_{z1}^k - c_{z2}^k \end{bmatrix}. \tag{4.4}$$

With Eq.(4.5) we can map any point from the target geometry $\mathbf{S}_T(0)$ to a point in $\mathbf{S}_S(0)$ represented in barycentric coordinates. We can see that this is not a one-to-one correspondence mapping, so it is a favorable circumstance of our method that works well with meshes with a different number of vertices or faces. $\varepsilon(T_{\mathscr{L}_j})$ represents the deformation of the triangle formed with landmarks $\mathscr{L}_j$ on $\mathbf{S}_S$ to adapt the geometry of the equivalent triangle in $\mathbf{S}_T$, and it is defined as $\varepsilon(\mathbf{T}_{\mathscr{L}_j}) = \mathbf{T}_{\mathscr{L}_j} - \mathbf{T}_{\mathscr{L}_j}^{tp}$.

Barycentric coordinates are also useful to compute a vector impact over a point inside a triangle when information is only available at the vertices, as in our case. So we take advantage of the current parametrization and obtain the displacement vector at the exact point with barycentric parametrization:

$$\mathscr{M} : \mathbf{v}_s^{i'} = \mathbf{T}_{\mathscr{L}_j^{-1}} \mathbf{T}_{C_S^k} \varepsilon \mathbf{T}_{\mathscr{L}_j^{-1}} (\mathbf{v}_t^i - \mathbf{t}_3 + \mathbf{c}_3^k + \mathbf{d})(-\mathbf{N}), \tag{4.5}$$

$$\mathscr{M}^{-1} : \mathbf{v}_t^i = \varepsilon \mathbf{T}_{\mathscr{L}_j} \mathbf{T}_{C_S^k}^{-1} (\mathbf{d}_i' \mathbf{N}) - \mathbf{c}_3^k - \mathbf{d} + \mathbf{t}_3). \tag{4.6}$$

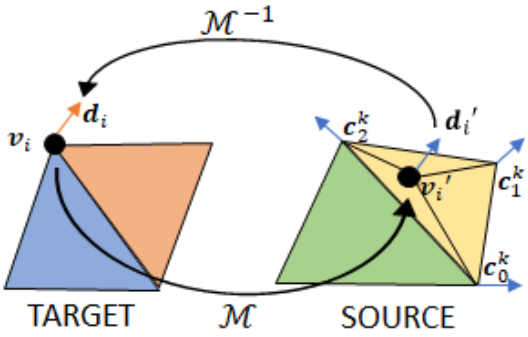


Figure 4.4 – **Graphical representation of the mapping and inverse mapping functions.** The different triangles represent distinct elements on source and target meshes. The illustration represents how a vertex on the target mesh $\mathbf{v}_i$ is represented in barycentric coordinates on the source mesh $\mathbf{v}'_i$. Then the displacements are computed according to the triangle vertices displacements $\mathbf{c}_1^k$, $\mathbf{c}_2^k$, and $\mathbf{c}_3^k$. $\mathbf{d}'_i$ represents the barycentric parametrization of the displacement vectors on the point $\mathbf{v}'_i$. $\mathbf{d}_i$ is the transferred displacement vector on the target point via the inverse mapping function.

Figure 4.4 depicts a graphical representation of the aforementioned terms.

4.6 Expression Transfer and Smoothing

4.6.1 Expression Transfer

With the displacement vectors computed at each of the vertices of the template, they can be transferred as a point using the inverse parametrization model described in Eq.(4.6). This mapping function allows returning the computed vectors to the initial target geometry $\mathbf{S}_T(0)$. So a point $v_R^i \in \mathbf{S}_T(P_b)$ can be computed as the addition of the mapped vector d_i to the initial point in $\mathbf{S}_T(0)$:

$$v_t^i(\mathscr{P}_b) = v_t^i(0) + d_i(\mathscr{P}_b)\,. \tag{4.7}$$

4.6.2 Smoothing energy

The central thought of the smoothing process is to provide smooth 3D expressions without missing detail. Given a transferred expression $\mathbf{S}_T(\alpha)$, the objective is to ob-

tain a smoothed shape $\mathbf{S}_T^S(\alpha)$ preserving as much detail as possible. The smoothing energy presented in Eq.(4.8) makes use of a specified data smoothing model (Total Variation) and a Gaussian connectivity weight over the expression vectors:

$$E = \sum_{l=1}^{L} \frac{1}{\omega_l \omega_d} \sum_{i \in V} \sum_{j \in N_l} \| s_t(\alpha)^i - s_t(\alpha)^j \|_2 \,. \tag{4.8}$$

Our approach minimizes the distance of a vertex regarding their neighbors for every point in the region of interest $\mathbf{S}_T(\alpha)$ using a Gaussian weighted Total Variation approach. The peculiarity of our smoothing functional is the use of a Gaussian weight over the l level neighbors (ω_l). Each vertex is considered a neighbor of another vertex just once, and this connection is done using the minimum value geodesic distance. Given the adjacency matrix, we can evaluate for each vertex of $\mathbf{S}_T(\alpha)$ its connectivities. So we can define, as neighbors of level l all of those neighbors with geodesic distance equals l which are nor included on a lower-level. So, for a vertex $\mathbf{S}_t(\alpha)^i$ with neighbors at level l, $\sum_{m=1}^{l-1} N_l \cap N_m = \varnothing$. The facial landmarks do not need to be estimated as they have a direct linear mapping, so they neither need to be smoothed.

4.7 Experimental Results

In this section, we evaluate the performance of the presented method using several datasets including both synthetic and real data. In Table 4.1 we report the used data, the proper citation and the number of vertices and triangular faces respectively. This data is relevant since offers a notion of the available vertices to transfer the expression wrinkles. Since, the most available vertices in the expression areas will produce more detailed expression wrinkles.

For quantitative evaluation, we provide the standard 3D error defined by:

$$\bar{e}_{3D} = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\mathbf{S}_{STS}(\alpha)^i - \mathbf{S}_S(\alpha)^i\|_F}{\|\mathbf{S}_S(\alpha)^i\|_F}, \tag{4.9}$$

where $\|.\|_F$ denotes the Frobenius norm. $\mathbf{S}_{STS}(\alpha)^i$ is the retransferred 3D expression over vertex i and $\mathbf{S}_S(\alpha)^i$ is the initial expression value over the same vertex. $\bar{e}_{3D}$ is computed over 3D shapes that have been previously aligned using Procrustes analysis.

We first evaluate our approach on a set of synthetic and real face reconstructions. For all of the datasets, we transfer a selected set of expressions from the dataset [39] (including surprise, kiss, pain, angry, sing, smile, and sad) to the initial reconstructions. Figure4.5 shows the performance of our method. In general terms,

Dataset	Synthetic/Real	Vertices	Faces	Wrinkles	Expression Wrinkles
Seq3 [36]	Synthetic	28,887	57,552	No	No
Ogre [5]	Synthetic	19,985	39,856	Yes	No
Mocap [4]	Synthetic	2,494	4,339	No	No
Face [36]	Real	28,332	56,516	Yes	No
Face1 [39]	Real	196,446	391,642	Yes	No
Face2 [39]	Real	196,446	391,614	Yes	No
Victor [39]	Synthetic	5,792	10,221	Yes	Yes

Table 4.1 – **Dataset resolution and level of detail.** For every employed dataset, we indicate the number of vertices and triangular faces, respectively. We also consider if the denoted datasets consist of synthetic sequences or real face videos. Last, we denote the level of detail included in the different inputs.

Dataset	Surprise	Kiss	Pain	Angry	Sing	Smile	Sad	$\bar{e}_{3D}$
Seq3 [36]	7.81	2.50	6.32	1.21	3.82	3.28	1.76	3.81
Mocap [4]	6.30	2.53	3.79	1.43	2.80	2.46	1.42	2.96
Ogre [5]	6.55	2.93	3.82	1.20	5.30	2.46	1.27	3.36
Face [36]	7.01	4.56	5.32	4.42	4.92	5.03	4.40	5.09
$\bar{e}_{3D}$	6.92	3.13	4.81	2.07	4.21	3.31	2.21	3.81

Table 4.2 – **Quantitative evaluation of synthetic and real datasets.** Percentage of 3D error over seven types of expressions on four datasets. Two types of analysis: average error per expression and per dataset.

our algorithm achieves even the most challenging of them. It should be noted as our approach can transfer the facial expression with different mesh resolutions, different facial geometries, and with the presence of a large change in the face topology, for instance, the single-eyed face of a cyclops.

Since we have no access to ground truth to obtain a direct measure of the standard 3D error, we have no other choice but to transfer our expression twice and compare the final result with the initial one. In particular, we first transfer the expression from the source to the target at rest, and second from the transferred expression to the source at rest. Then the standard 3D error can be properly computed using the initial expression of the source model as ground truth. The major inconvenience with this procedure is that expression transferring is estimated twice, so the error is going to be greater.

We quantitatively evaluate our method with the before mentioned datasets and expressions and obtain the 3D mean error described in Table4.2. Our mean 3D

Figure 4.5 – **Qualitative evaluation of face expression transfer for four different datasets.** In all cases, we display a neutral shape, together with seven expressions denoted surprise, kiss, pain, angry, sing, smile, and sad, respectively. **First row:** A projected view of the 3D source model. **From second to fifth row:** Our 3D estimation in a frontal view of the datasets: Seq3, Mocap, Ogre, and Face, respectively. Observe that our expression transfer algorithm produces very accurate solutions for all cases, even when it is applied for complex shapes like the Ogre dataset. It is worth noting that our approach obtains also nice results for noisy shapes, such as the Face sequence.

error across the different datasets and challenging expression is 3.8%. However, it is worth noting that all the quantitative results are produced by a double transferring process, then errors may be increased.

We graphically interpret the previous analysis in Fig.4.7. In this figure, we represent the 3D error located over the face with Face sequence used as the target. We apply the double transferring as before to detect the areas that produce more inaccuracies. We represent the corresponding error per vertex for seven expressions. In general terms, the errors are a product of the facial detail inconsistencies in high-frequency areas, for instance, a winkle was transferred on the result but it is not present on the original 3D.

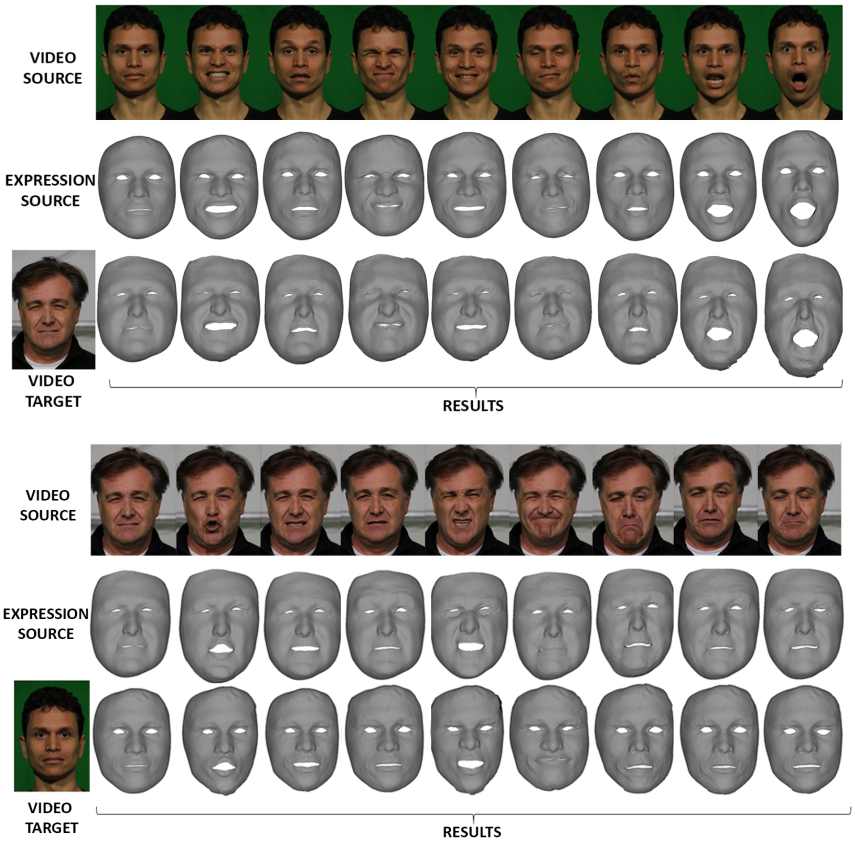


Figure 4.6 – **Qualitative evaluation of on-line facial expression transfer.** The same information is displayed in both cases. **Top and Middle:** Source image frames and the corresponding 3D estimation. **Bottom:** The left-most picture displays our target face model. We also represent our 3D estimation after the facial expression transfer, considering the source facial gesture in the top row. In all cases, our approach produces high detailed solutions in 3D where original wrinkles and folds are preserved and new expression wrinkles are transferred.

The second experiment consists of two on-line facial expressions transferring sequences with real data. With this purpose, we consider two datasets with very

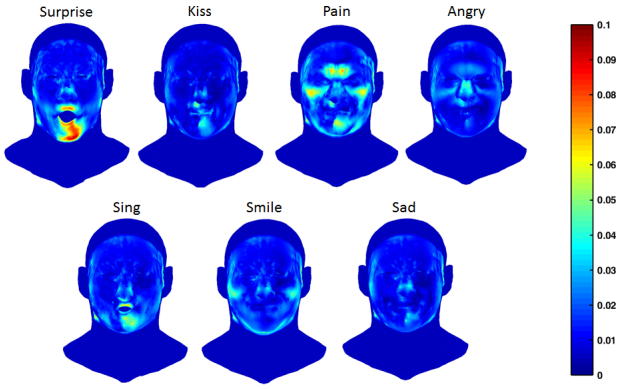


Figure 4.7 – **Source-target-source transfer.** Once an arbitrary facial expression is transferred from the source to the target, we transfer back from target to source. We display the distribution of the 3D errors for seven facial primitives, considering the Face [36] sequence to define the target model. Bluish areas mean the double transfer is more accurate.

dense data from a video (Face 1 and Face 2). In the first part, we consider the face in Face 1 video as our shape at rest and extract the expressions from Face 2 video. We apply our algorithm to transfer the expressions from Face 2 to our 3D target face extracted from Face 1 video. In the second part, we consider the inverse procedure. We consider the face in Face 2 video as our shape target and transfer the expressions from Face 1 video. A qualitative evaluation of this experiment is presented in Figure4.6. The top figure relates the first part and the bottom figure, the second part. Figure4.6 displays the effectiveness of our method transferring expressions obtained from all the frames in the video, even transferring and preserving the high-frequency details. For both experiments our conclusion is similar, the presented results manifest the generality of our method as well as the effectiveness in the transferring of the expressions and details.

4.8 Conclusion

In this chapter, we have discussed our solution to address the problem of automatically 2D-to-3D transfer of facial expressions. With this purpose, we presented our approach, which is fully unsupervised. It includes a sub-region mapping model parametrized with barycentric coordinates. These features convert our method

in a robust and efficient technique, which works in a dense setup and can handle topology changes, different mesh resolutions, and noise in data. Moreover, our method can properly work without any training data at all, and it does not consider any unreachable expression to transfer.

During this later chapter, we have extensively validated our approach over different challenging facial expressions of both synthetic and real data. We have shown that it has superior performances over all the proposed datasets yielding in realistic expressions and preserving the fine detail of the input face. Our method also can transfer expression wrinkles and folds from the given template if there is enough resolution.

5 A Complete Pipeline to Generate Detailed Expresive Characters from a Single RGB Image

In this chapter, we present the unification of the different procedures exposed in this thesis as a novel framework for detailed 3D reconstruction of human faces, including a set of blending shapes. Our method works from a single RGB image and unknown illumination conditions. The skin wrinkles and facial hairs are retrieved via texture analysis and modeled through new parametrization models. We validate our method to precisely describe different skin details and facial hairstyles with a simple number of parameters. We then produce a set of subject-specific blending shapes via facial mesh-to-mesh mapping. Based on these correspondences, we estimate the 3D displacement vectors on the reconstructed facial geometry from a given template. As described in the previous chapters, our method supports accurate shape recovery as facial details and facial hairstyles, ranging from wrinkles and scars to elaborate beard styles. It can perform with several mesh resolutions and partial occlusions. We demonstrate the effectiveness of the full pipeline over several datasets.

5.1 Introduction

In the previous chapters, we introduced different techniques to improve the quality of facial reconstructions and expression transferring along with detail. Our first proposal was an effective and efficient wrinkle detection and modeling procedure, departing from a single RGB image and a volumetric regressed initial face. Later on, we described our facial hair parametric model, in addition to the necessary detection and reconstruction methodology to detect and fit the parameters to adjust the single image input. Lastly, we defined our expression transfer model that can relocate expression from a coarse initial face, or rigged template to a new facial mesh, for instance, the detailed reconstructed facial mesh in the previous chapters.

In the first chapter, we have described our method to effectively recover wrinkles and face detail from a single RGB image and unknown illumination parameters. We have demonstrated that our approach can effectively retrieve different types of

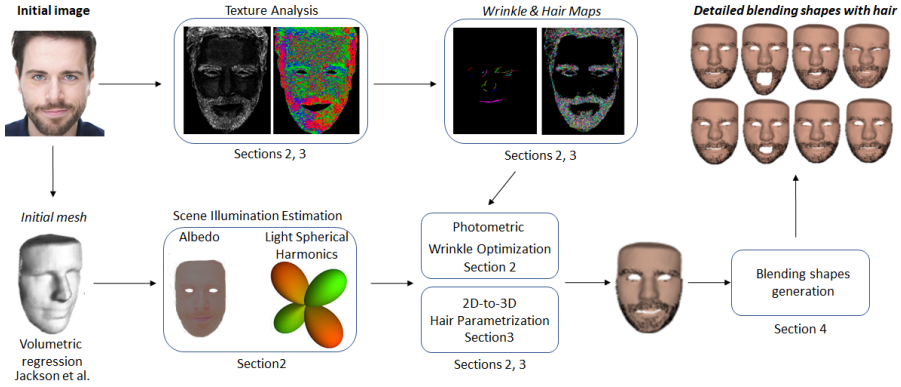


Figure 5.1 – **Main stages of the full pipeline** including detail recover (described in Chapter 2), facial hair recovery (depicted in Chapter 3), and automatic blending shape generation (exposed in Chapter 4).

detail from medium to high-frequency detail from human faces producing realiable solutions compared with the single-view state of the art methods. The benefits of our approach are not only the resulting reconstructions but the fact that it runs with the commodity of a laptop, and it does not require training data.

In Chapter 2, we proposed a new parametric approach to recover facial hair from a single RGB image. Our method can effectively deal with the complexity of the 3D hair structure. We have shown our approach to describe hair strands in 2D with texture information extracted from a single RGB image. We proved how we model the detected hair fibers in 3D using a parametric hair model based on 3D helixes, later on, we completed our approach with an effective energy-based method to obtain the hair parameters from the 2D described hair strands.

Our energy-based hair model can work in the same conditions as the previous wrinkle model, not requiring expensive hardware nor training data, and requiring the computation of exclusively a small number of parameters. For this reason, we found it interesting to connect both approaches to produce more realistic results with both facial wrinkles and hair.

The aftermost approach described in this dissertation differs from the first and second since it is not a reconstruction model based on a single image, but it is strongly related to detail. In the third chapter, we defined our expression transfer model based on a subregion mapping approach and a barycentric coordinate system. It aims to transfer expressions between different facial meshes preserving detail as well as transferring new expression wrinkles and microexpressions from

Algorithm 4 Algorithm including the main steps of facial esptression transfer.

INITIALIZATION
Coarse initialization with [51] $\mathbf{I} \rightarrow \mathbf{V}$
Volume to mesh conversion $\mathbf{V} \rightarrow \mathbf{S}$
Scene illumination estimation $\mathbf{I} = \rho \cdot (\mathbf{l} \cdot \mathbf{Y}(\mathbf{n}(\mathbf{S})))$

WRINKLE RECONSTRUCTION
Wrinkle detection $\mathbf{I}(x,y)^k$
Detail map construction DT by applying Eq. (2.8)
Depth estimation d via Eq. (2.10)
Depth transferring $\mathbf{S}_i^{detailed} = \mathbf{S}_i^{coarse} + \mathbf{d}_i$

FACIAL HAIR RECOVRERY
Texture Analysis $\mathbf{I} \rightarrow (\mathbf{M},\mathbf{O})$
Hair detection from texture $(\mathbf{M},\mathbf{O}) \rightarrow \mathbf{H}$
Hair tracing and endpoint labeling $\mathbf{H} \rightarrow \mathbf{P}$
Hair clustering $\mathbf{P} \rightarrow \mathbf{P_k}$
Energy minimization $\boldsymbol{F}_i^h(\boldsymbol{p}^h, \boldsymbol{n}^h, l, w, r, \theta, g)$
Appending further realism $\boldsymbol{F}_i^h(\boldsymbol{p}^h, \boldsymbol{n}^h + \lambda_o, l + \lambda_l, w, r, \theta, g)$

EXPRESSION TRANSFER
Obtain facial landmarks $\mathscr{L}$ via [8]
Mesh aligning with landmarks $\mathbf{S}_S(0), \mathbf{S}_T(0)$
Low-resolution correspondences $\mathscr{T}(\mathscr{L})$ with Eq. (4.2)
KNN of unasigned vertices
Super-resolution correspondences $C_S(\mathscr{L})$ with Eq. (4.2)
Mapping of expression vectors $\mathscr{M}$ and $\mathscr{M}^{-1}$ with Eq (4.5) and (4.6)
Transferring the expression vectors with Eq. (4.7) Smoothing of the resultant vector set with Eq. (4.8)
Moving the hair root position acording to new coordinates $\boldsymbol{F}^h(\boldsymbol{p}^h + \Delta\boldsymbol{p}, \boldsymbol{n}^h, l, w, r, \theta, g)$
Recompute triangle normals and correct hair orientations $\boldsymbol{F}^h(\boldsymbol{p}^h, \boldsymbol{n}^h + \Delta\boldsymbol{n}, l, w, r, \theta, g)$

a template. As in the previous approaches, we do not need complex hardware nor training data. Notwithstanding, in this case, we require a rigged reference containing the expression vectors to transfer.

In this chapter, we connected the different methods presented in this disser-

tation in a combined output. This joint pipeline offers not only a detailed reconstruction of a human face, including wrinkles and facial hair plus the automatic generation of different facial expressions. In summary, the construction of an entirely rigged digital human from a single RGB image. A schematic outline was depicted in Figure 5.1 and in Algorithm 4.

5.2 Detailed reconstruction

We depart with a single image from which we extract the facial landmarks [8] and the volume [51], as is described in Chapter 2. Later on, we adapt the regressed volume to a functional mesh by reclaiming the exterior part **v**. Later on, we iterate with the image and the coarse mesh to obtain the illumination parameters **l** and ρ_i. Subsequently, we estimated the wrinkle positions in 2D with Gabor orientation responses and attain the wrinkle description through two distances dP and dC. Finally, we obtain the corresponding depth **d** by studying the photometry of the scene. As a final result, we achieve a detailed facial surface in just a few seconds.

5.3 Hair recovery and the effect of orientation correction

The second step of the pipeline has its detailed description in Chapter 3. As a summary, we grab the high-frequency detail face computed in Chapter 2 along with the wrinkle map and estimate the hair textures of the facial image **H**, as an adaption, we eliminate from **H** all the blobs that fully correspond with a blob in the wrinkle map $\mathbf{w}^k$. Later, we trace the hairs in ordered sets of pixels from root to tip, considering hair crossings and self-intersections $\mathbf{p}^h$.

We estimate the parameters defining the 3D structure of the hair with the proposed method and append further detail. It is worth noting that the best detailed the face is, the most changes in the hair orientations happen, and the most realist the final result is.

5.4 Wrinkle preservation on expresion and appending of further expression wrinkles

To preserve skin wrinkles and combine with expression furrows, there are two crucial factors to consider. The first is the smoothing procedures used during the expression transfer process can not over smooth or remove detail on the input face.

Dataset	Row 1	Row2	Row3	Row4	Row5	Row6
Image Reso.	849×1273	1156×980	1024×768	1760×2640	1750×1168	1920×1088
Wrinkles	33	27	4	32	39	25
Time (s)	3.270798	3.133740	2.188152	3.018402	3.320594	2.978279
[51](all)	0.25054	0.23941	0.25260	0.30128	0.32763	0.32445
[51](wrink.)	0.24591	0.24910	0.24029	0.30760	0.40851	0.29763
Ours (all)	0.24961	0.23717	0.24953	0.30127	0.32225	0.3129
Ours (wrink.)	0.23583	0.23574	0.23827	0.29389	0.40723	0.29617
Hairs Up.	766(61)	72(8)	62(2)	66(9)	334(67)	377(20)
Hairs Lo.	10251(510)	610(60)	9575(155)	7733(440)	9384(361)	0(0)

Table 5.1 – **Final Numeric Evaluation.** Quantitative evaluation of the different steps of the pipeline over the pictures displayed in Fig.5.4. This table depicts the resolution of the input image, the number of detected wrinkles as well as the number of recovered hairs in the upper and lower face. The numbers in parenthesis represent the synthesized hairs.

Luckily, our smoothing therm presented in Chapter 4 accomplish this criterion. Second, is that wrinkles appended due to expression must be added to consider the current detail and integrate both the expression and skin details in the current features. Since our method work on barycentric coordinates, it guarantees that all the points lying on the same triangle will have coherent and refined vector transfer, including the points with high-detail.

5.5 Hair animation on expression

Animating the facial hair according to the facial expression is a challenging task due to the hair structural complexity. However, with our parametric model, we can represent a hair with its root position, orientation, and a set of 5 different parameters. This simple representation allows us to efficiently change the position and the orientation of the hair strand.

First, we find out the barycentric coordinates of each hair root employing the input-output function described in Equation 4.2. Then with the new coordinates, we apply the inverse mapping function described in Equation 4.6 to move the full hair strand to its new position.

Last, we estimate the difference in the orientation of the initial and final triangles

representing the facial surface and reorientate the hair according to this difference.

$$\boldsymbol{F}^h(\boldsymbol{p}^h + \Delta\boldsymbol{p}, \boldsymbol{n}^h + \Delta\boldsymbol{n}, l, w, r, \theta, g), \tag{5.1}$$

where $\Delta\boldsymbol{p}$ represents the root displacement after animation and $\Delta\boldsymbol{n}$ represents the root orientation change after animation.

5.6 Experimental Results

Within the course of this dissertation, we have proposed different alternatives to gather more realistic facial reconstructions from a single RGB image input. Moreover, we have validated each of the presented methods with a great set of different examples. In this chapter, we have moved beyond, introducing a joint framework to recover human faces with great detail and with its corresponding blending shapes.

Starting with qualitative results, Figure 5.2 shows the partial results of detailed reconstruction and har acquisition from a single RGB image. Since it is not trivial to find data that contains strong facial features with facial hair we also provide in Figure 5.3 a close-up figure in which a particular case can be observed in more detail. Finally, in Figure 5.4 we present our final joint results of detailed faces with hair and their corresponding blending shapes generated from a template face via expression transfer.

We can appreciate compelling results in each of the steps individually as has previously demonstrated in this dissertation, but we also achieve interesting results when the entire parts are working collectively. We can recover different types of small facial detail, moreover, the expression transfer holds the reconstruction detail and append further expression wrinkles when it is necessary (for example, see cheek detail in smile expression).

Quantitative metrics support the previous results. We disclose the number of wrinkles, the energy minimization residual, and the number of hairs Table5.1. It can be observed how our method slightly improves the solution provided by [51] in a numerical comparison in detail recovery. However, due to the small variation in the 3D mesh produced by wrinkles, the improvement can be better observed in the qualitative analysis.

5.6.1 Failure Cases and Limitations

Since our approach lies principally on image derivatives, there is a group of conditions that may be detrimental to the final result. These circumstances encompass severe shadows, texture-varying areas, obscure tattoos, significant occlusions due to thick beards or large orientation changes. Regardless of the outstanding perfor-

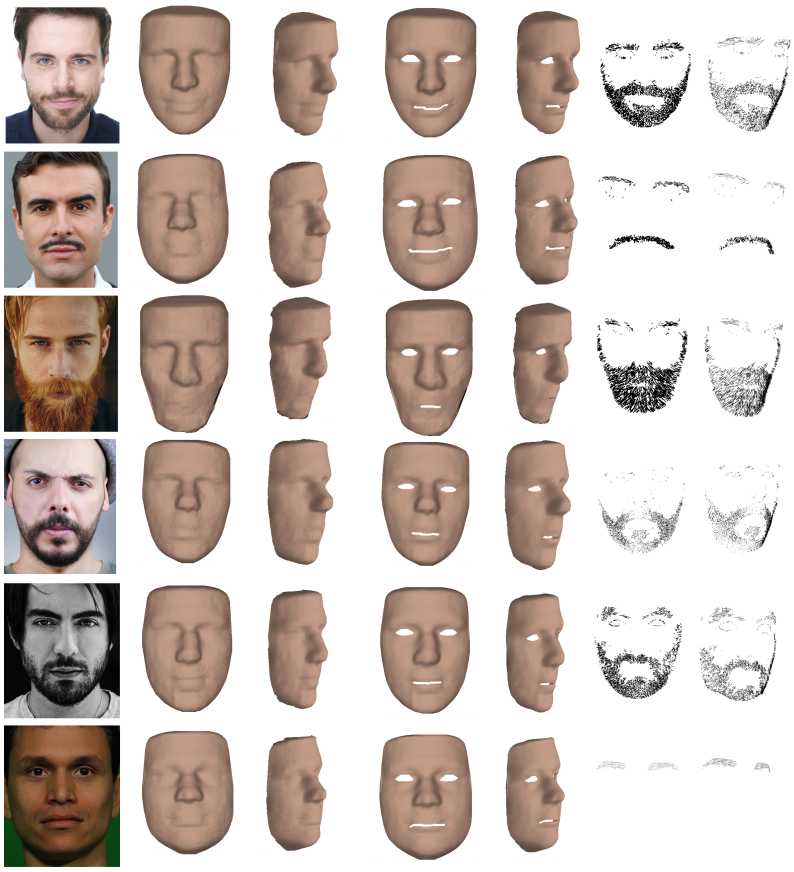


Figure 5.2 – **Comparative of detail recovery on the different datasets employed in this chapter. First column:** Original RGB image, **Second and third rows:** Reconstrucion of [51] with different view points, which is also our initializations. **Fourth and fifth rows:** Our detailed reconstruction from different viewpoints. **Seventh and eight rows:** The reconstructed hair from different viewpoints (front and side).

mance of our approach in general situations, the previous scenarios may result in inaccurate results.

Figure 5.3 – **Comparative of detail recovery on a closeup example around the eye area. First:** Original RGB image, **Second:** Reconstrucion of [51], which is also our initializations. **Third:** Our detailed reconstruction.

5.6.2 Discussion

Facial detail and hair fibers are regularly defined via mere textures or alpha maps. In this dissertation, we presented a method to successfully acquire the geometry of these elements, which allows a joint computation and representation of the facial model. Our approach describes a reliable, fast, and effective alternative to the previous methods. To the best of our knowledge, it is the first commitment covering not only the detail and facial hair recovery in a joint manner but also to accomplish an automatic creation of blending shapes from the retrieved data.

5.7 Conclusion

We have presented a novel framework to automatically reconstruct detailed faces and produce their blending shapes from a single RGB image without any training data. It joints the detailed reconstruction from a single RGB image, including wrinkles and facial hair, with the automatic generation of blending shapes for animation purposes. It represents an advance in the topic, considering it tackles in the same pipeline three of the big problems in detailed reconstruction and animation, and it requires only a single RGB image of a human face. Our method detects and models different facial details through an energy minimization solution. Additionally, it generates the corresponding blending shapes for the reconstructed faces via direct mapping, which allows us to transfer any expressions to the reconstructed face, not limiting to a prebuild basis set. With this setup, we can efficiently process images to recover ready-to-animate 3D characters with a considerable level of detail from solely a single RGB image. We have evaluated the setup with different images and datasets to provide both qualitative and quantitative analysis. We demonstrated

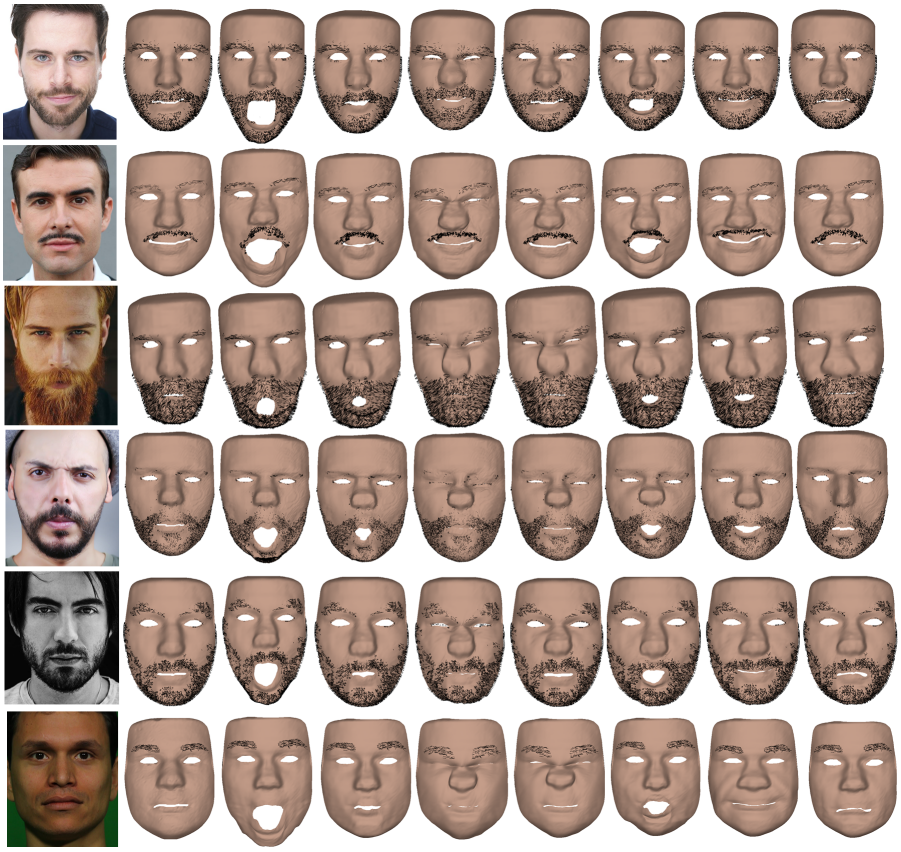


Figure 5.4 – **Qualitative evaluation of our full pipeline on real images in the wild.** Each row represents a different dataset and each column a different generated expression. First column displays the input image. Second column displays the 3D neutral recovered expression with hair. From third to eighth column are provided the generated blending shapes expressions of Surprise, Kiss, Pain, Angry, Sing, Smile and Sad.

the ability of our model to deal with several types of wrinkles and facial hairstyles, as well as the capability to produce smooth blending shapes from different data sources. The results are encouraging and present the opportunity to generate 3D data for different purposes, including generating synthetic data with sight in deep learning techniques.

Clausula Part III

Lifelike rendering of Siren by Epic Games artists [7].

6 Conclusions and Future work

6.1 Conclusions

In this Ph.D. dissertation, we have addressed the problem of recovering and dealing with detailed reconstruction from different perspectives. In the first chapter, we presented a comprehensive introduction to the problem of recovering and dealing with high-frequency detail on human faces. The first part of the presented methods englobes the procedures to recover the high-frequency detail. In the second chapter, we addressed the problem of recovering wrinkle detail and other high-frequency detail on facial skin. In the third chapter, we proposed a model to effectively recover facial hair. Both approaches are solved from single RGB image as input. In the second part, we introduced a novel method to expression transfer and blending shape generation which can generate new expressions preserving original detail and appending new expression wrinkles.

First, we have proposed an intuitive and effective approach to retrieve 3D detailed reconstruction of faces from a single RGB image.Currently, state-of-the-art techniques working with deep-learning require large amounts of annotated data, which in practice is laborious to obtain. In particular, wrinkles and facial hairs are hard to achieve since they demand to be annotated individually by hand. Still, it is insufficient to recover personal detail accurately. While our method considers only an RGB image as an input and does not require any training data at all. In an unsupervised setup, it can obtain several features to parametrize the wrinkles even without having been observed previously. This scheme allows us to model not only regular wrinkles but, person-specific attributes, such as scars or several shapes as a consequence of aging. Our approach is fast and efficient, retrieving person-specific details in few seconds by sorting out a reduced domain photometric optimization problem. We have extensively evaluated our approach on several datasets including synthetic and real images, considering a wide range of variability in which we have outperformed existing state-of-the-art solutions. An interesting avenue for future research is to extend our formulation to handle more severe occlusions as well as a validation in real-time at frame-rate.

In Chapter 3, we addressed the problem of recovering 3D facial hair from a single RGB image without any training data. We first had proposed a parametric 3D hair model based on a 3D helix and a set of energies that rely on 2D detections over different facial areas. With our method, we directly estimate some features over the input image in 2D and extract the 3D structure as a fitting problem of our parametric model. As in the previous chapter, it does not require any training data, user interaction, or any specific setup. In both cases, we have broadly endorsed our proposal over a collection of several images with uncontrolled illumination and show consistent and realistic results even in challenging cases as thick beards, eyeglasses, low-resolution pictures, and eyebrow makeup.

Relating to hair fiber reconstruction experiments, we had compared our approach with the current state-of-the-art method [13] and despite the clear disadvantage in terms of hardware of our single-image versus their multiview setup, our method retrieves very competing results.

In Chapter 4, we have presented our solution to address the problem of automatically 2D-to-3D transfer of facial expressions. With the guidance of the described sub-region mapping model, with barycentric coordinates, we develop a fully unsupervised model that does not require any training data at all. Our model has been proved to be robust and efficient even handling dense and detailed faces at any mesh resolution, even with noise in data. Since our method does not rely on a low-rank established from a training dataset, there is no unreachable expression to transfer. We just need a reference model with the given expression to reproduce it on our data. We have extensively validated our model over a large set of data including both synthetic and real. For all the datasets, we tested the most challenging expressions. We have shown that it has superior performances over all the proposed datasets yielding in realistic expressions and preserving the fine detail of the input face. Further, our method can transfer expression wrinkles and detail from the given template if there is enough resolution to reproduce it.

In Chapter 5, we introduced a joint approach in which we aim to reconstruct detailed faces and provide their blending shapes from a single RGB image without any training data. It results in detailed human faces with well-appointed details including wrinkles, wounds, scars, facial hair, and expression details. It requires only an RGB image of a face and no further information, user interaction, or training data. Additionally, we generate different expressions on ready-to-animate 3D detailed characters. We validated our approach over several data and pointed out the efficiency and coherence of the final results, validating our approach for different purposes such as generating 3D synthetic 2D data with sight in deep-learning techniques.

6.2 Future Perspective

In this thesis, we studied the reconstruction and animation of detailed faces from the perspective of model-based approaches represented with geometrical and physical priors. The principal strengths of the presented method are their low requirements in terms of data, computation power, and time budget. By studying the 2D properties and estimating the 3D component via optimization methods, we have achieved a certain level of detail in the skin and facial hair, which allows the character to own specific features. These accurate details are essential to represent humanlike digital characters, as well they are helpful to provide them a personal story and a personality, for instance, scars and wounds or with specific facial hairstyles.

On the other hand, we believe that deep-learning on human faces is a research field with a bright future: a lot of studies are done yearly in several areas from surveillance, face reenactment, reconstruction, animation, etc. However, the need for enormous amounts of data is imperative for these methods to obtain reliable results. The obstacle a researcher in these areas faces the most is the lack of annotated data required by supervised models, which is a laborious task. Considering faces are usually protected personal data, it is challenging to find real datasets that can provide the required volume of faces with enough detail. One of the future lines of this thesis is to study the best path to generate synthetic faces through reconstructing faces from the wild and combining the results to synthesize new ones. It is decisive to develop new accurate faces to obtain detailed results, as well as generate sufficient expressions per subject to have a representative set.

Furthermore, we keen to explore other avenues, for instance, to investigate the wrinkle space with the focus on achieving better representations of facial detail holding the speed and time budget performance we have presented in this work. A case in point would be to explore not only solutions of higher-order but to provide the resultant faces with micro details as skin texture.

We apprehended that realistic facial hair is likewise a solid source of realism to represent a digital person. One of our future work lines consists of improving the naturalness of our model by developing hair density and crossings with credible representations. We especially wish to sharpen our model on eyebrows and eyelashes since these parts are the most contributing not only in 3D reconstruction but on later facial animations. Besides, we aim to recover facial hair on other head parts out of the face like the neck or the whiskers.

Finally, moving to the animation side, we believe that build a set of micro-expressions joint with detailed reconstructions is a fundamental part of the accomplishment of humanlike characters. We aim to extend our expression transfer model to focus on the blinking wrinkles area and into the effect of eyebrow hair animation

on expressions.

6.3 Scientific Articles

This dissertation has led to the following communications:

6.3.1 Published Journals

- **Gemma Rotger**, Francesc Moreno-Noguer, Felipe Lumbreras and Antonio Agudo. Detailed 3D Face Reconstruction from a Single RGB Image. Journal of the World Society of Computer Graphics, Vol 27, Num 2, 2019.

6.3.2 Submitted Journals

- **Gemma Rotger**, Francesc Moreno-Noguer, Felipe Lumbreras and Antonio Agudo. Animating detailed 3D faces with hair from images. Computer Vision and Image Understanding. **Under Review**.

6.3.3 International Conferences and Workshops

- **Gemma Rotger**, Felipe Lumbreras, Francesc Moreno-Noguer and Antonio Agudo. 2D-to-3D Facial Expression Transfer. International Conference on Pattern Recognition (ICPR), Beijing, (China), 2018. (**Poster**).

- **Gemma Rotger**, Francesc Moreno-Noguer, Felipe Lumbreras and Antonio Agudo. Detailed 3D Face Reconstruction from a Single RGB Image. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen (Czech Republic), 2019. (**Oral**).

- **Gemma Rotger**, Francesc Moreno-Noguer, Felipe Lumbreras and Antonio Agudo. Single View Facial Hair 3D Reconstruction. Iberian Conference on Pattern Recognition and Image Analysis, Madrid (Spain), 2019. (**Oral**).

Bibliography

[1] A. Agudo, J. Montiel, B. Calvo, and F. Moreno-Noguer. Mode-shape interpretation: Re-thinking modal space for recovering deformable shapes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 03 2016.

[2] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Modal space: A physics-based model for sequential estimation of time-varying shape from monocular video. *Journal of Mathematical Imaging and Vision (JMIV)*, 57(1):75–98, 2017.

[3] A. Agudo and F. Moreno-Noguer. Learning shape, motion and elastic models in force space. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[4] A. Agudo and F. Moreno-Noguer. Global model with local interpretation for dynamic shape reconstruction. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[5] A. Agudo and F. Moreno-Noguer. A scalable, efficient, and accurate solution to non-rigid structure from motion. *Computer Vision and Image Understanding (CVIU)*, 167(2):121–133, 2018.

[6] A. Agudo and F. Moreno-Noguer. Shape basis interpretation for monocular deformable 3D reconstruction. *IEEE Transactions on Multimedia*, PP:1–1, 09 2018.

[7] E. Alvarez. With 'Siren,' Unreal Engine blurs the line between CGI and reality. https://www.engadget.com/2018-03-22-siren-epic-games-unreal-engine-vicon.html. Accessed: 2020-07-19.

[8] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[9] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(4), 2017.

[10] Y. Bando, T. Kuratate, and T. Nishita. A simple method for modeling wrinkles on human skin. In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 166–175, 2002.

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

[12] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (TOG)*, volume 29, page 40, 2010.

[13] T. Beeler, B. Bickel, G. Noris, P. Beardsley, S. Marschner, Robert W. Sumner, and M. Gross. Coupled 3D reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics (TOG)*, 31(4):117, 2012.

[14] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross. Multi-scale capture of facial geometry and motion. In *ACM Transactions on Graphics (TOG)*, volume 26, page 33, 2007.

[15] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM SIGGRAPH*, 1999.

[16] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):40, 2013.

[17] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. In *ACM Transactions on Graphics (TOG)*, volume 29, page 41, 2010.

[18] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG)*, 34(4):46, 2015.

[19] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014.

[20] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3D shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013.

[21] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2014.

[22] J. Cao, Y. Li, and Z. Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] X. Cao, Z. Chen, A. Chen, X. Chen, S. Li, and J. Yu. Sparse photometric 3D face reconstruction guided by morphable models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] M. Chai, L. Luo, K. Sunkavalli, N. Carr, S. Hadap, and K. Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (TOG)*, 34(6):204, 2015.

[25] M. Chai, T. Shao, H. Wu, Y. Weng, and K. Zhou. Autohair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

[26] M. Chai, L. Wang, Y. Weng, X. Jin, and K. Zhou. Dynamic hair manipulation in images and videos. *ACM Transactions on Graphics (TOG)*, 32(4):75, 2013.

[27] Y. Chen, Z. Song, S: Lin, R. R. Martin, and Z.Q. Cheng. Capture of hair geometry using white structured light. *Computer-Aided Design*, 96:31–41, 2018.

[28] Z. Chen, Y. Ji, M. Zhou, S. B. Kang, and J. Yu. 3D face reconstruction using color photometric stereo with uncalibrated near point lights. *arXiv preprint arXiv:1904.02605*, 2019.

[29] L. P. Chew. Constrained delaunay triangulations. *Algorithmica*, 4(1-4):97–108, 1989.

[30] N. Chinaev, A. Chigorin, and I. Laptev. Mobileface: 3D face reconstruction with efficient cnn regression. In *European Conference on Computer Vision (ECCV)*, 2018.

[31] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *arxiv preprint arXiv:1711.09020*, 2017.

[32] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[33] C. Dong. Modelling of captain John Price, Modern Warfare 4. https://www.artstation.com/artwork/QAa1d. Accessed: 2020-07-19.

[34] G. Fyffe, A. Jones, O. Alexander, R. Ichikari, and P. Debevec. Driving high-resolution facial scans with video performance capture. *ACM Transactions on Graphics (TOG)*, 34(1):1–14, 2014.

[35] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, S. Jones, H. Li, and P. Debevec. Multi-view stereo on consistent face topology. In *Computer Graphics Forum*, volume 36, pages 295–309. Wiley Online Library, 2017.

[36] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[37] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *International Journal of Computer Vision (IJCV)*, 104(3):286–314, 2013.

[38] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen, P. Perez, and C. Theobalt. Automatic face reenactment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[39] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics (TOG)*, 32(6):158–1, 2013.

[40] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016.

[41] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)*, 30(6), 2011.

[42] M. Gleicher. Animation from observation: Motion capture and motion editing. *ACM SIGGRAPH Computer Graphics*, 33(4):51–54, 1999.

[43] P. F. U. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3D reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[44] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 55–66. ACM, 1998.

[45] E. Guillou, D. Meneveaux, E. Maisel, and K. Bouatouch. Using vanishing points for camera calibration and coarse 3D reconstruction from a single image. *The Visual Computer*, 16:396–410, 11 2000.

[46] H. Hachmann, M. Awiszus, and B. Rosenhahn. 3D braid guide hair reconstruction using electroluminescent wires. *The Visual Computer*, pages 1–12, 2018.

[47] G.S.J. Hsu, Y.L. Liu, H.C. Peng, and P.X. Wu. RGB-D-based face reconstruction and recognition. *IEEE Transactions on Information Forensics and Security*, 9(12):2110–2118, 2014.

[48] L. Hu, C. Ma, L. Luo, and H. Li. Robust hair capture using simulated examples. *ACM Transactions on Graphics (TOG)*, 33(4):126, 2014.

[49] L Hu, C. Ma, L. Luo, and H. Li. Single-view hair modeling using a hairstyle database. *ACM Transactions on Graphics (TOG)*, 34(4):125, 2015.

[50] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3D morphable face model and fitting framework. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2016.

[51] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[52] W. Jakob, J. T. Moon, and S. Marschner. Capturing hair assemblies fiber by fiber. *ACM Transactions on Graphics (TOG)*, 28(5):164, 2009.

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

[54] Y.C. Lee, J. Chen, C.W. Tseng, and S.H. Lai. Accurate and robust face recognition from RGB-D images with a deep learning approach. In *British Machine Vision Conference (BMVC)*, volume 1, page 3, 2016.

[55] J. Li, W. Xu, Z. Cheng, K. Xu, and R. Klein. Lightweight wrinkle synthesis for 3D facial modeling and animation. *Computer-Aided Design*, 58:117–122, 2015.

[56] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3D face reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016.

[57] S. Liu, Z. Wang, X. Yang, and J. Zhang. Realtime dynamic 3D facial reconstruction for monocular video in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[58] G. Longhi et al. Second chance games and visual effects, inc. https://www.artstation.com/artwork/zWma2. Accessed: 2020-07-19.

[59] L. Luo, H. Li, and S. Rusinkiewicz. Structure-aware hair capture. *ACM Transactions on Graphics (TOG)*, 32(4):76, 2013.

[60] F. Lupo. Digital double from 3D scan and some expressions. http://frankinolupo.blogspot.com/2012/07/digital-double-from-3d-scan-and-some.html. Accessed: 2020-07-19.

[61] S. Paris, H. M. Briceño, and F.X. Sillion. Capture of hair geometry from multiple images. *ACM Transactions on Graphics (TOG)*, 23(3):712–719, 2004.

[62] S. Paris, W. Chang, O. I. Kozhushnyan, W. Jarosz, W. Matusik, M. Zwicker, and F. Durand. Hair photobooth: geometric and photometric acquisition of real hairstyles. In *ACM Transactions on Graphics (TOG)*, volume 27, page 30. ACM, 2008.

[63] J. Peissig, T. Goode, and P. Smith. The role of eyebrows in face recognition: With, without, and different. *Journal of Vision (JOV)*, 9(8):554–554, 2009.

[64] F. Pighin, R. Szeliski, and D. H. Salesin. Resynthesizing facial animation through 3D model-based tracking. In *IEEE International Conference on Computer Vision (ICCV)*, 1999.

[65] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. GANimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128, 08 2019.

[66] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, 2016.

[67] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[68] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[69] G. Rotger, F. Lumbreras, F. Moreno-Noguer, and Antonio A. 2D-to-3D facial expression transfer. In *International Conference on Pattern Recognition (ICPR)*, pages 2008–2013, 08 2018.

[70] F. Shi, H.T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):222, 2014.

[71] P. Sirio. The evolution of Nathaniel Drake in Uncharted game series. https://www.gamepur.com/. Accessed: 2020-07-19.

[72] R. W. Sumner and J. Popović. Deformation transfer for triangle meshes. *ACM Transactions on Graphics (TOG)*, 23(3):399–405, 2004.

[73] Y. Sun, J. Dong, M. Jian, and L. Qi. Fast 3D face reconstruction based on uncalibrated photometric stereo. *Multimedia Tools and Applications*, 74(11):3635–3650, 2015.

[74] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *European Conference on Computer Vision (ECCV)*, 2014.

[75] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[76] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6):183–1, 2015.

[77] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[78] L. Tian, J. Liu, and W. Guo. Three-dimensional face reconstruction using multi-view-based bilinear model. *Sensors*, 19(3):459, 2019.

[79] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)*, 9(2):137–154, 1992.

[80] A. Tuan Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[81] L. Valgaerts, C. Wu, A. Bruhn, H.P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Transactions on Graphics (TOG)*, 31(6):187–1, 2012.

[82] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433, 2005.

[83] Y. Wang, C.C.L. Wang, and M.M.F. Yuen. Fast energy-based surface wrinkle modeling. *Computers & Graphics*, 30(1):111–125, 2006.

[84] Y. Wei, E. Ofek, L. Quan, and H.Y. Shum. Modeling hair from multiple views. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 816–820. ACM, 2005.

[85] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. In *ACM Transactions on Graphics (TOG)*, volume 30, page 77, 2011.

[86] M. Wilczkowiak, E. Boyer, and P. Sturm. Camera calibration and 3D reconstruction from single images using parallelepipeds. In *IEEE International Conference on Computer Vision (ICCV)*, 2001.

[87] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics (TOG)*, 32(6):1–11, 2013.

[88] X. Yu, Z. Yu, X. Chen, and J. Yu. A hybrid image-cad based system for modeling realistic hairstyles. In *SIGGRAPH*, 2014.

[89] M. Zhang, M. Chai, H. Wu, H. Yang, and K. Zhou. A datadriven approach to four-view image-based hair modeling. *ACM Transactions on Graphics (TOG)*, 36(4):156, 2017.

[90] Y. Zhou, L. Hu, J. Xing, W. Chen, H.W. Kung, X Tong, and H. Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2018.

[91] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, et al. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

[92] M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb. State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum (Eurographics State of the Art Reports 2018)*, 37(2), 2018.

[93] M. Zollhöfer, J. Thies, D. Bradley, P. Garrido, T. Beeler, P. Péerez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37:523–550, 05 2018.