

Catarina Alexandre Nunes Runa Miranda

# Real-time Motion Capture Facial Animation



Departamento de Ciência de Computadores  
Faculdade de Ciências da Universidade do Porto  
November 2015

Catarina Alexandre Nunes Runa Miranda

# Real-time Motion Capture Facial Animation



*Tese submetida Faculdade de Ciências da  
Universidade do Porto para obtenção do grau de Doutor  
em Ciência de Computadores*

Departamento de Ciência de Computadores  
Faculdade de Ciências da Universidade do Porto  
November 2015



**To my Family...**

*Climb mountains not so the World can see you,  
but so you can see the World.  
David McCullough Jr.*

# Acknowledgments

I started to climb this mountain few years ago. After working at industry (healthcare sector), I was quite lost searching for a way to change my path and follow my dream of learning more about graphics and computer vision in an open and challenging environment. Then, I met Verónica and her amazing team at Porto Interactive Center (PIC). It was love at first glance and for four years I had the opportunity of being part of it. I cannot be more grateful than I am right now. Without the help of every single person at PIC, I would not be able to learn so much and be always so motivated to finish this PhD thesis.

First, I have to express my deepest gratitude to my advisor, Verónica Orvalho. She was the one that trusted me and had the patience to teach me all that I know today. Verónica was more than an advisor, she was my biggest source of inspiration giving always the best advices in my everyday struggle. She challenged me and she brought the best of me every single day.

I also deeply thank to Xenxo Alvarez, who helped and gave me all the technical support and knowledge. On the other hand, Xenxo was a friend and an anchor when I lost track of my research and life, rescuing me from the big monsters of CG.

Pedro Mendes. My best friend and the best programmer I know. Without you, Pedro, nothing of this would be possible. You have been always there, as a friend and as a work partner. Working with you is a daily pleasure and I hope to help you in your PhD journey the way you helped me. I always will be here for you.

PICantes. The incredible, the powerful and best team ever. Without you guys, nothing of this would be possible and I will be eternally grateful. Special thanks to José Serra, colleague and friend, always there even when I was doing crazy stuffs. Also, Marta Meira, for the patience during the long reviews and for the wise advices.

Miguel Sales Dias and João Freitas from Microsoft for their hospitality and partnership during my time at Microsoft and during these years of research.

I also thank to UPIN team, especially to Filipe Castro, for mentoring me during business development using PIC work, including my PhD research.

I am also very grateful to Professor Miguel Coimbra and Professor Carlos Salema from Instituto de Telecomunicações for giving me all the support and dedication to help me finish my PhD graduation.

My friends and family Runa, who showed an endless comprehension and support, giving me the strength to go on. Even in this last year of isolation to write my thesis, you were always there.

My beloved brother, his wife and my sweet nieces. César, I am so thankful for all you have done for me and, even in my stupid mistakes, you never gave up on me. I love you. Andreia, Matilte and Carolina, you were always in my mind being a voice inside saying come on, you can do it!.

Last, but the most, my parents. The best parents in the world. For their unconditional love and support. Without you I would not be who I am and I would not achieved any of this. I love you more than anything and I will be forever grateful for all the sacrifices and efforts you have done for me. Mom, for being simply there, anytime, always hiding the tears for not being with you. You are the strongest woman I know, teaching me if you fall 7 times, you have to get up 8. Father, you were the one that make me this "starving for knowledge and progress" person. You also teach me to work really really hard to achieve my goals and the best skill ever: be brave and never give up. Father, you always will be my first source of inspiration and the man of my life.

# Abstract

Humans rely on facial expressions for communication since ancient times. Looking through the evolution of telecommunications, we clearly observe an increase of the demand for face-to-face interactions. General users started by using phones and computers. But, the massive usage of these technologies was seen when image and video were introduced (e.g. videoconferencing, share photographs, etc). Today, recent advances in Virtual Reality (VR) at consumer-level (Oculus VR 2014) lead to a "tipping-point" in communication and human-computer interactions (HCI). Now, it becomes possible to enter a virtual environment (e.g. films, videogames, interactive applications) and communicate with other people through a virtual avatar, represented by a 3D character. However, expert artists follow the user's morphology and behaviors to create the 3D characters' unique shape, appearance and movements, making this process time-consuming and labor intensive.

*This PhD thesis presents an automatic markerless facial Motion Capture (MoCap) system that maps the unique facial data of a person into a 3D character. A traditional facial animation pipeline is usually divided in three stages: modeling, rigging animation. This thesis focuses in the animation process. Driven by the entertainment industry tendency, MoCap technologies are adopted to accelerate the creation of high quality facial animation. But MoCap remains challenging to non-expert users, since the capture of facial movements rely on expensive and complex hardware setups and/or require user-dependent tedious calibrations.*

This thesis Facial MoCap system is defined by a collection of modular methods that automate the facial animation process. Our methods run in real-time and use off-the-shelf hardware, like webcams, to fit consumer-level applications, including VR environments.

We explored the fundamental science behind MoCap systems, human perception and facial animation. We have conducted two studies to validate facial features, mor-

phologies and behaviors. Study one was responsible for exploring facial features' self-perception. Study two delivers a real-time method of facial features extraction and emotions recognition with an accuracy of 94%. As a result of these studies, we created three novel methods: (i) a real-time facial MoCap tracker, (ii) a mapping algorithm to map motion data to a 3D character and (iii) three VR methods for facial MoCap under partial occlusions produced by VR headsets.

We defined and created a new protocol and a proof of concept database called FdMiee to validate our facial MoCap methods. FdMiee's protocol simulates a wide range of environmental parameters (e.g. lighting variations) and facial behaviors (e.g. facial expressions). In addition, by adjusting testing parameters, this protocol can be used to validate generic Computer Vision (CV) algorithms. FdMiee is available to the public to use.

The methods and research developed in this thesis have been included, presented and validated in: the EU projects VERE - Virtual Embodiment and Robotic Re-Embodiment and GOLEM - Realistic Virtual Humans; the national project LIFEis-GAME - Learning of Facial Emotions using Serious Games.

With this thesis, we deliver a modular solution to create facial animation on-the-fly. As a result, we produce methods that can be used *by anyone for everyone* contributing to the next-generation of human communication and interaction through the combination of virtual environments and MoCap.

# Resumo

Desde a pré-história que os seres humanos recorrem a expressões faciais para comunicar. Observando a evolução das telecomunicações é evidente a procura por uma comunicação cara-a-cara. O público comum começou por usar telefones e computadores. Mas a utilização massiva destas tecnologias foi verificada quando introduzimos a transmissão de imagem e vídeo (por exemplo, chamadas por videoconferência, partilha de fotografias). Nos dias de hoje, com os mais recentes avanços em Realidade Virtual (RV) ao nível do consumidor comum (Oculus VR 2014), encontramos-nos num "ponto de viragem" das comunicações e interações homem-máquina (HCI). Agora, é possível entrar no mundo virtual (e.g. filmes, videojogos, aplicações interactivas) e comunicar com outras pessoas através de um *avatar*, isto é, uma personagem 3D.

Contudo, para que estes personagens 3D estejam preparados para reproduzir a morfologia e estrutura dos movimentos faciais do utilizador, artistas com elevado grau de expertise têm de executar um processo manual bastante moroso e intensivo.

*Esta tese de doutoramento apresenta um sistema automático e sem marcadores para captura de movimentos faciais (MoCap) que mapeia os movimentos característicos do utilizador para uma personagem 3D.* As linhas de produção de animação facial tradicionais dividem-se, em geral, em três estágios: modelação, *rigging* e animação. Tendo em conta as tendências da indústria do entretenimento, os sistemas de MoCap têm sido adoptados de forma a acelerar a criação de animação facial de elevada qualidade. Mas estes sistemas não são adequados para o público comum, visto que a captura de movimentos faciais é executada com recurso a equipamento caro e complexo e/ou requerendo calibrações manuais e demoradas por parte do utilizador.

O sistema de MoCap facial apresentado por esta tese é definido por um conjunto de métodos modulares que automatizam o processo de animação facial. Os nossos métodos correm em tempo-real e usam equipamento acessível ao público comum, tais como *webcams*, para aplicações interactivas, onde se incluem os ambientes RV.

Exploramos a ciência fundamental inerente aos sistemas de MoCap, percepção humana e animação facial. Conduzimos dois estudos para validar os pontos característicos, morfologias e movimentos faciais. O primeiro estudo foi responsável por explorar a nossa percepção dos pontos característicos da nossa própria face. O segundo estudo criou um método de extração de movimentos faciais e reconhecimento de emoções em tempo-real com uma precisão de 94%. Como resultado deste estudos, criámos três métodos inovadores: (i) um *tracker* MoCap facial em tempo-real, (ii) um algoritmo de mapeamento que transfere os movimentos capturados para uma personagem 3D, e (iii) três métodos de RV para MoCap facial compatível com oclusões criadas pelo equipamento de RV.

Definimos um protocolo e adquirimos uma database, nomeada FdMiee, para validar os nossos métodos de MoCap facial. O protocolo FdMiee simula e captura diferentes condições ambientais (por exemplo, variações de luminosidade) e movimentos faciais de participantes (por exemplo, expressões faciais). Adicionalmente, através do ajuste destes parâmetros ambientais e faciais, o protocolo adapta-se e pode ser usado para validação de algoritmos de visão por computador em geral. FdMiee está publicamente disponível.

Os métodos investigados e desenvolvidos durante esta tese de doutoramento foram incluídos, apresentados e validados em: projectos EU, VERE - *Virtual Embodiment and Robotic Re-Embodiment* e GOLEM - *Realistic Virtual Humans*; e um projecto FCT, LIFEisGAME - Learning of Facial Emotions using Serious Games.

Com esta tese, desenvolvemos soluções modulares para criação de animação facial em tempo-real. Como resultado, produzimos métodos que podem ser usados *por todos e para todos* contribuindo para a próxima geração de comunicações e interações humanas através da combinação de ambientes virtuais e MoCap.



# Contents

<b>Abstract</b>	<b>7</b>
<b>Resumo</b>	<b>9</b>
<b>List of Tables</b>	<b>18</b>
<b>List of Figures</b>	<b>26</b>
<b>1 Introduction</b>	<b>27</b>
1.1 Motivation and Impact . . . . .	28
1.2 Scientific challenges . . . . .	30
1.3 MoCap Framework . . . . .	32
1.4 Contributions . . . . .	34
1.5 Application Domain . . . . .	36
1.5.1 Entertainment Industry . . . . .	36
1.5.2 Healthcare Sector . . . . .	38
1.5.3 Educational Interactive Systems . . . . .	38
1.6 Outline . . . . .	39
<b>2 MoCap Fundamental Science</b>	<b>41</b>
2.1 Face Image Task: Self-perception of Facial Features . . . . .	42

2.1.1	Background . . . . .	43
2.1.2	Methodology . . . . .	44
2.1.3	Results . . . . .	44
2.1.4	Discussion and Conclusions . . . . .	49
2.2	Real-Time Emotion Recognition . . . . .	51
2.2.1	Background . . . . .	51
2.2.2	Methodology:	
	Geometric Features Extraction Method . . . . .	53
2.2.2.1	Eccentricity Features . . . . .	55
2.2.2.2	Linear features . . . . .	56
2.2.2.3	Extracted Features . . . . .	58
2.2.3	Classifier validation . . . . .	58
2.2.3.1	Features Evaluation . . . . .	59
2.2.4	Discussion and Conclusions . . . . .	62
<b>3</b>	<b>Facial MoCap Tracking</b>	<b>65</b>
3.1	Background . . . . .	65
3.2	Methodology . . . . .	69
3.2.1	Novel Method: Facial MoCap Tracking . . . . .	69
3.2.1.1	Features Definition . . . . .	69
3.2.1.2	Method's Overview . . . . .	69
3.2.1.3	Calibration . . . . .	72
3.2.1.4	Runtime . . . . .	74
3.2.1.5	Reset . . . . .	78
3.3	FdMiee's protocol . . . . .	78
3.3.1	FDMiee: Protocol Methodology . . . . .	79

3.3.2	FDMiee: Definition and Validation . . . . .	81
3.3.2.1	Requirements . . . . .	82
3.3.2.2	Acquisition Hardware . . . . .	82
3.3.2.3	Environment-Change Generation Equipment . . . . .	82
3.3.2.4	Protocol Guidelines . . . . .	83
3.3.3	Protocol Validation . . . . .	84
3.3.4	Protocol Conclusions . . . . .	84
3.4	Results and Discussion . . . . .	86
3.5	Conclusions . . . . .	92
<b>4</b>	<b>MoCap Facial Animation</b>	<b>95</b>
4.1	Background . . . . .	96
4.1.1	Traditional Facial Animation . . . . .	96
4.1.2	Facial Animation with MoCap . . . . .	98
4.2	Methodology . . . . .	100
4.2.1	Calibration . . . . .	101
4.2.2	Runtime . . . . .	103
4.2.2.1	Geometric Mapping Algorithm . . . . .	105
4.2.2.2	Creating Animation . . . . .	107
4.3	Results and Discussion . . . . .	110
4.4	Conclusions . . . . .	116
<b>5</b>	<b>MoCap VR Methods</b>	<b>117</b>
5.1	Background . . . . .	118
5.1.1	Persistent Partial Occlusions: a today's problem . . . . .	119
5.1.2	Partial Occlusions and Expressiveness . . . . .	120

5.2	Methodology . . . . .	122
5.2.1	VR Persistent Partial Occlusions: a novel method . . . . .	123
5.2.2	VR Assessing Facial Expressions . . . . .	125
5.2.2.1	VR Emotion Recognition: novel method . . . . .	127
5.2.2.2	VR Facial Expressions Predictor: novel method . . . . .	128
5.3	Results and Validation . . . . .	129
5.3.1	MoCap VR method: Persistent Partial Occlusions . . . . .	129
5.3.2	MoCap VR method: Assessing Facial Expressions . . . . .	129
5.3.2.1	MoCap VR Emotion Recognition . . . . .	132
5.3.2.2	MoCap VR Facial Expressions Predictor . . . . .	133
5.3.2.3	MoCap VR Assessing Facial Expressions: Visual Results	135
5.4	Conclusions . . . . .	138
<b>6</b>	<b>Conclusions and Future Directions</b>	<b>141</b>
6.1	Conclusions . . . . .	141
6.2	Future Directions . . . . .	144
6.3	Take Home Message . . . . .	145
<b>A</b>	<b>Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation</b>	<b>147</b>
<b>B</b>	<b>Real-Time Emotion Recognition: a Novel Method for Geometrical Facial Features Extraction</b>	<b>161</b>
<b>C</b>	<b>VERE Poster: Facial Tracking Systems</b>	<b>171</b>
<b>D</b>	<b>Facial Expressions Tracking and Recognition: Database Protocols for Systems Validation and Evaluation</b>	<b>173</b>

E	VERE Poster: Real-time facial animation through motion capture	193
F	Assessing Facial Expressions in Virtual Reality Environments	195
G	Facial emotions as a metric of 21st Century competencies: a draft protocol for acquiring and classifying facial emotion data for learning analytics	209
	References	219



# List of Tables

2.1	Average localization errors for each feature in cm. . . . .	46
2.2	Factor scores for horizontal X and vertical Y components. . . . .	47
2.3	The subset of anthropometric facial landmarks used to calculate our proposed geometric facial features. . . . .	54
2.4	The eight ellipses used to extract the eccentricity features (for the landmark labels please refer to Figure 2.3). . . . .	57
3.1	Protocol's flexible and fixed variables. . . . .	80
3.2	MoCap Tracking method - Percentage of failure in face detection under different environment changes. Results comparison between (i) Saragih <i>et al.</i> [SLC11a], (ii) Our facial MoCap Tracker and (iii) Cao <i>et al</i> [CHZ14].	87
3.3	MoCap Tracking method - Percentage of failure in detection subtle and macro expressions of different participants (i) Saragih <i>et al.</i> [SLC11a], (ii) Our facial MoCap Tracker and (iii) Cao <i>et al</i> [CHZ14]. . . . .	89
5.1	k-Fold CRM Accuracy comparison to scenario (i) and to the scenario (ii). Results in percentage (%). . . . .	132
5.2	Statistical Analysis of scenario (i) - Results in percentage (%). . . . .	133
5.3	Statistical Analysis of scenario (ii) - Results in percentage (%). . . . .	133
5.4	k-Fold CRM Accuracy comparison facial expressions assessed (Eye-brows Up or Down) with subset <i>S1</i> and <i>S2</i> . Results in percentage (%). . . . .	134

5.5	Eyebrow Up prediction - Statistical Analysis to subsets $S1$ . Results in percentage (%).	134
5.6	Eyebrow Down prediction - Statistical Analysis to subsets $S1$ . Results in percentage (%).	135



# List of Figures

1.1	VERE VR environment using 3D characters example generated using our in-house animation system. . . . .	28
1.2	LIFEisGAME game mode: the card shows which emotion the user has to express to win the game. Then, with the capture of the user's face using a webcam, our algorithms recognize the emotion and transfer user's facial movements to the 3D character. . . . .	29
1.3	MoCap Framework: the three colored boxes represent the three research stages of this PhD thesis. The white boxes contain the methods delivered, which are distributed as pre-processes or runtime processes. .	32
1.4	VR MoCap Framework: the purple box contains the VR MoCap methods that replace the facial MoCap Tracking in the main MoCap Framework. . . . .	34
1.5	Broadcast application: Gollum face in <i>The Hobbit: The Battle of The Five Armies (2014)</i> by Weta Digital Ltd. . . . .	37
1.6	Real-time application: Wither's face at <i>The Witcher 3: Wild Hunt (2015)</i> by CD Project Red. . . . .	37
2.1	Biases in face representation: A: Schematic of feature locations used to instruct participants. B: Actual and mean represented locations. C: Average of 50 female faces reproduced with permission from Perception-Lab [Per]. Blue arrows indicate mean judgment error for each feature. D: Average female face adjusted according to the mean represented locations of our participants. E, F: as for C, D with average of 50 male faces. . . . .	45

2.2	Association between horizontal and vertical distortion factors demonstrates variation in representation of face shape across individuals. Results of a canonical correlation between the horizontal (X1,X2) and vertical (Y1,Y2) factors. A: Vectors showing the principal feature loadings ( $>0.4$ or $<-0.4$ ) of the factors, adjusted by the coefficients indicating important ( $>0.4$ or $<-0.4$ ) contributions to the canonical variate. The vector lengths are shown at 4x the actual values for visual clarity. Note the negative sign for Y1 coefficient. B: Average female and male faces implied by a low and high score on the canonical variate. Note that the canonical variate separates long and thin from short and wide face representations. . . . .	48
2.3	The subset composed by 19 points of the 66 facial landmarks used to extract our proposed geometric facial features. . . . .	55
2.4	The definition of the first (a.), " <i>upper</i> " and the second (b.), " <i>lower</i> " ellipses of the mouth region using respectively the triple $(A_M, B_M, U_{m1})$ and $(A_M, B_M, D_{m2})$ . . . . .	56
2.5	The final results of the eight ellipse construction (a). Eccentricities of the facial ellipses changes according to the person's facial emotion (b). .	57
2.6	Results using a Random Forests classifier for each dataset composed by a sub-set of features of a subset of emotions to classify. * means without considering contemptuous emotion, ** without considering neutral emotion, *** without considering neutral and contemptuous emotions . . . . .	60
2.7	Accuracy comparison of emotion facial recognition methods(not differential or differential features) with six universal emotions. . . . .	60
2.8	Confusion matrix with Random Forest using all eight emotions for subsets $S3 - S4 - S5$ . . . . .	60
2.9	Confusion matrix with Random Forest using 6 emotions (without neutral and contemptuous) for subsets $S3 - S4 - S5$ . . . . .	61
3.1	Marker-based MoCap example: Actor Mark Ruffalo in the role of Hulk at " <i>The Avengers: Age Of Ultron</i> " (2015). Source: <a href="http://www.cosmicbooknews.com">http://www.cosmicbooknews.com</a>	68

3.2	Markerless video-based approaches - tracking comparison (from left to right): (i) User-specific algorithm [CWLZ13], (ii) 3D CLM [SLC11b] and (iii) DDE user-free algorithm [CHZ14]. . . . .	68
3.3	Hierarchy structure containing facial zones and landmarks used to setup the XML configuration file. . . . .	71
3.4	Zones and landmarks hierarchy example: The blue box is type Zone vertical, id 1 and position order. This Zone 1 has as childs two landmarks (type), X1 and X2 with id 1 and 2, respectively. The purple box is the Zone id 2, horizontal and contains the child landmarks X3, X4 and X5 with id 3, 4 and 5, respectively. . . . .	72
3.5	Novel method for facial MoCap Tracking: inputs in diamond boxes, method's stages (colored boxes) and respective sub-methods. . . . .	73
3.6	MoCap Tracking method - Calibration stage that receives the image stream and XML configuration file and returns the ROI, BME value and Hierarchy structure of the user. . . . .	73
3.7	Face model landmarks of facetracker of Saragih <i>et al.</i> [SLC11a, SLC11b].	74
3.8	MoCap Tracking method - Hierarchy structure template and parameters required. In the template XMin, XMax, YMin and YMax define the zone limits, where the maximum between the min and max of the coordinates is used to build the zone ellipse. . . . .	75
3.9	MoCap Tracking method - Visual result of hierarchy structure: landmarks belong to zones (i.e.ellipses) with the same color. . . . .	75
3.10	MoCap Tracking method - Runtime stage uses the outputs from Calibration and the image stream and run the Optical Flow combined with stabilization methods to update de 2D landmarks stored in hierarchy format and the ROI. This stage also contains a Failure Check. . . . .	76
3.11	MoCap Tracking method - Optical Flow result, before (left image) and after (right image) BME filtering. . . . .	76

3.12 MoCap Tracking method - Landmark update using Optical Flow: the Optical Flow displacement that influence the position of the landmark X (green area) is defined by the interception between zone limits (blue ellipse) and circumference with radius defined by influence radius (black circle). . . . .	77
3.13 Failure detection when hierarchy structure of the facial model is not maintained: at time 1 we have the landmarks X1, X2 and X3 in the correct order. A wrong update at runtime lead to a change of X1 and X2 positions. Failure detection method detects the error and reverts the positions of the landmarks X1 and X2 at time2 to their correct positions at time1. . . . .	78
3.14 MoCap Tracking method - Failure detection resets the face model using the image stream and updates the ROI and Hierarchy structure. . . . .	79
3.15 FDMiee samples results for HD Camera (A), Webcam (B) and IR Camera (C) . . . . .	85
3.16 MoCap Tracking method - Environment light variation - Low (left) and High (right) conditions. . . . .	88
3.17 MoCap Tracking method - Environment Background variation - White (left), Static features (middle) and Dynamic features (right). . . . .	88
3.18 MoCap method - Environment with Multiperson - Static (left) and Dynamic person (right). . . . .	88
3.19 MoCap Tracking method - Behavior parameters: Head moving: Yaw (left), Pitch (middle) and Roll (right). . . . .	89
3.20 MoCap Tracking method - Behavior parameters: Expressions: Disgust (left) and Joy (right). Red arrows represent the face model tracking and green points to the Optical Flow. . . . .	90
3.21 Comparison - Behavior parameters: Expressions: Disgust detected by Saragih <i>et al.</i> [SLC11a] (left) and Cao <i>et al.</i> [CHZ14] (right). . . . .	90
3.22 Cao <i>et al.</i> method [CHZ14] - Environment light variation - Low light with failure (left) and High light with no failure (right) detection. . . . .	91

3.23	Cao <i>et al.</i> method [CHZ14] - Environment with dynamic Multiperson scenario - test case with correct tracking (left), but after moving, the tracker fails detection (middle) or the detection jumps to the secondary participant (right). . . . .	91
3.24	Cao <i>et al.</i> method [CHZ14] - Examples of failure in tracker's detection during the change of facial behaviors - Head moving: Yaw (left), Pitch (middle) and Roll (right). . . . .	92
3.25	Saragih <i>et al.</i> method [SLC11a] - Examples of detection failure during the change of facial behaviors. . . . .	92
4.1	Blendshapes' rig example: different poses of upper part of the face [SSMCP02]. . . . .	97
4.2	Bone-based rig example with Unreal game engine. ( <i>Copyright 2001-2007 Epic Games</i> ) . . . . .	98
4.3	Rig data-flow structure. Stages: (i) requirements; (ii) input motion to activate the rig; (iii) geometry deformation. [OBP <sup>+</sup> 12] . . . . .	99
4.4	Lightstage - acquisition hardware used in Digital Ira [DHT <sup>+</sup> 00, AFB <sup>+</sup> 13, vdPJD <sup>+</sup> 14]. . . . .	100
4.5	Digital Ira results [AFB <sup>+</sup> 13, vdPJD <sup>+</sup> 14]. . . . .	101
4.6	Mapping method overview scheme (from left to right): inputs (MoCap and Rig parameters), calibration and runtime stages. . . . .	102
4.7	Mapping method - calibration stage receives MoCap and 3D character parameters and uses the algorithms at rectangular boxes to return a T,R,S Space Calibration transform (Global and Local) and an Hashtable containing the correspondence between the landmarks and the vertex of the 3D character's mesh. . . . .	103
4.8	Mapping method - Raycast method to calculate the third coordinate of 2D landmarks. Red dots are the input 2D landmarks and the white dots are the resultant 3D landmarks. . . . .	104
4.9	Mapping method - Colored landmarks that user selects in the 3D character. . . . .	104

4.10	Mapping method - Runtime stage that uses the 2D landmarks stream from the MoCap tracking system, the outputs from calibration and 3D character (mesh and rig) to calculate 2D joints movements that are used in the Animation stage. . . . .	105
4.11	Mapping method - Animation Space definition, where the blue dot is a 2D joint. . . . .	108
4.12	Mapping method - Animation Runtime scheme that uses rig information and the 2D joints movements calculated in runtime to produce 3D character's animation. . . . .	108
4.13	Mapping method - Animation weight calculation example, where the red dot 2D translation of the joint of control. . . . .	109
4.14	Mapping method - Animation Curve example. . . . .	110
4.15	Mapping results - Surprise facial expression reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time. . . . .	111
4.16	Mapping results - Surprise facial expression with jaw asymmetrical movement reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time. . . . .	111
4.17	Mapping results - Surprise facial expression with jaw asymmetrical movement reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time. . . . .	112
4.18	Mapping results - Example of inner lip undesirable effect in the 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time. . . . .	112
4.19	Mapping results - Example of joy-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time. . . . .	113
4.20	Mapping results - Example of disgust-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time. . . . .	114
4.21	Mapping results - Example of sadness-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time. . . . .	114

4.22	Mapping results - Example of surprise-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time. . . . .	115
4.23	Mapping results - Example of speech sequence animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.	115
5.1	VR hardware-based setup proposed by Li <i>et al.</i> [LTO <sup>+</sup> 15] to overcome partial occlusions issue. . . . .	119
5.2	Examples of diversity of facial expressions created by mixing two basic emotions [McC06]. . . . .	121
5.3	MoCap VR methods' framework: filled blue and purple boxes refer to our VR methods distributed as pre-processes or runtime processes. . . .	123
5.4	MoCap VR hardware setup. . . . .	124
5.5	VR setup examples with: nVisor SX111 (left) and Oculus Rift DK2(right).	124
5.6	MoCap VR method: Persistent partial occlusions. From left to right: calibration image without VR HMD; our method uses cut point (red circle) to cut image an overlay at subsequent images: at left, what facial MoCap method see is a full face and, at right, the real image. . .	125
5.7	MoCap VR methods: Expressions predictor training (purple) and emotion predictor training (blue) with CK+ database. . . . .	127
5.8	MoCap VR method results: Persistent Partial Occlusions method applied to Saragih <i>et al.</i> [SLC11a] MoCap. The real image (left), our method result and what MoCap processes (middle) and final result from our method (right). . . . .	130
5.9	MoCap VR method results: Persistent Partial Occlusions method applied to Cao <i>et al.</i> [CHZ14] MoCap. . . . .	130
5.10	MoCap VR method results: Persistent Partial Occlusions method applied to general occlusion created by a paper <i>et al.</i> [CHZ14] MoCap. . .	131
5.11	VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (green). Check that our emotion and prediction match original image eyebrows movements (green box). . . .	136

5.12 VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (red). Check that our emotion and prediction match original image eyebrows movements (green box). . . .	137
5.13 VR Assessing Facial Expressions: Correct Emotion Recognition result (blue) and no Expression Predictor result, since there is not movement. Check original image in green box. . . . .	137
5.14 MoCap VR Assessing Facial Expressions: Incorrect Emotion Recognition result (blue) and Expression Predictor result. Check original image to see that Expression Predictor is correct (green box). . . . .	138



# Chapter 1

## Introduction

*Human's communication relies in facial expressions to reflect cognitive status, emotions and intentions not provided by speech. Since early age, we are educated and trained to recognize and perform body and facial gestures, learning how to connect them to specific emotions, intentions, situations, experiences and consequences. These human skills make us experts in facial analysis and in pin pointing abnormal movements in other's face. Therefore, reproducing human facial movements in virtual characters is a challenging task for both animators and artists and almost impossible for non-expert users. This makes facial animation one of the most expensive processes within a 3D characters production pipeline. Motion capture (MoCap) systems are widely used during the animation process to reduce experts manual work and increase believability of the results during the animation process. However, MoCap technologies require costly acquisition hardware or usage of intrusive marker-based systems. On the other hand, cheaper and markerless solutions rely on complex calibrations not suitable to non-experts and do not support persistent partial occlusions. Solving these problems become a need with the recent advances of Virtual Reality (VR) hardware at consumer-level by Oculus VR, where on-the-fly facial animation of 3D characters is essential for non-verbal communication. The overall goal of this PhD thesis is to deliver modular methods for markerless MoCap facial animation using off-the-shelf hardware, reducing time and costs. This chapter introduces the background, motivation and challenges that lead to the PhD research. Then, we overview our MoCap framework and highlight scientific contributions retrieved.*



Figure 1.1: VERE VR environment using 3D characters example generated using our in-house animation system.

## 1.1 Motivation and Impact

This PhD research started in 2012, while working at VERE European (EU) and LIFEisGAME Portuguese projects. VERE EU project focuses in the creation of immersive virtual environments, dissolving the boundaries between the human body and surrogate representations in virtual reality and physical reality. Our contribution to VERE consisted in the development of an application where the user is able to see himself in a mirror as a 3D character (see Figure 1.1). To deliver the desired environment immersion, the 3D character needed to reproduce the user's unique facial movements. In the LIFEisGAME project, the main goal was to teach facial emotions to ASD (Autism Spectrum Disorder) children through videogames and state of the art Computer Vision (CV) and Computer Graphics (CG) technologies. So, to LIFEisGAME, we proposed a game mode where children had to make facial movements to match the expressions asked by the game. In this game mode, children see their facial movements mirrored in the game's 3D character, improving the learning of expressions through self-evaluation (see Figure 1.2).

Thus, the projects had distinct goals but both required a common solution for real-time markerless facial animation of 3D characters using off-the-shelf hardware. The main technical difference between VERE and LIFEisGAME is that in VERE half of the face is occluded, as the participant is wearing an Head Mounted Display (HMD), while



Figure 1.2: LIFEisGAME game mode: the card shows which emotion the user has to express to win the game. Then, with the capture of the user’s face using a webcam, our algorithms recognize the emotion and transfer user’s facial movements to the 3D character.

in LIFEisGAME all the face is visible. To achieve high quality MoCap the system must have the following features: real-time, markerless and use off-shelf-hardware. It was necessary to develop a real-time and markerless solution, because: (i) the solution’s users are non-experts in CG or CV, like therapists, which are not able to place properly the markers; (ii) markers are too intrusive and uncomfortable to general users, particularly to children and (iii) 3D character should mimic user’s facial movements in real-time to allow user’s fully embodiment, thus achieving the illusion of presence [SSV14]. In addition, the solution should use off-the-shelf hardware, like webcams, hence being usable in therapeutic and schools environments, where it is impossible to setup or buy complex acquisition hardware for facial capture.

Today, there is a great demand for MoCap facial animation of 3D characters for communication and interaction with virtual environments [BMW<sup>+</sup>06]. For proper human communication, 3D characters need to be prepared to transmit a synergistic combination of verbal (e.g. speech or text) and non-verbal (e.g. speech, facial expressions, gestures, signs) signals. Non-verbal signals enrich verbal communication giving additional information of interlocutors’ cognitive status, emotions and intentions. Consequently, non-verbal communication, such as facial expressions, is fundamental for human social interactions [JJ13]. The evolution of telecommunications and interactive digital media also enhances the necessity to transmit both signals, searching for

solutions that better reproduce face-to-face communication. At the beginning, long distance telecommunications, like phone and computer-based communications, were only able to transmit speech or text. But soon, technological developments lead to the introduction of additional communication channels, like images and video stream using webcams [Bio97]. Current revolution of global communications makes possible the interlocutors' interaction within virtual environments (e.g. films, videogames, interactive applications). This "tipping-point" become explicit with the recent availability of Virtual Reality (VR) headsets at consumer-level (Oculus VR 2014). But again, we are missing the non-verbal communication channel, i.e. we are able to transmit speech but virtual characters facial animation on-the-fly is still in early stage.

Despite the technological efforts made by the academy and industry in the last decade, facial animation still appears as a bottleneck in animation pipelines relying in a considerable amount of human input [Lew06, vdPJD<sup>+</sup>14, LTO<sup>+</sup>15]. In traditional pipelines, animators and artists execute a labor-intensive and manual work to reproduce human's face movements and generate convincing facial animation [OBP<sup>+</sup>12]. Diversity of facial traits and characters visual styles [Sco93] enhance these difficulties. As a solution, MoCap methodologies have been widely used. MoCap provides real face movements input to trigger animation avoiding their manual generation from the scratch. However, MoCap methodologies are not suitable for non-experts users, because they require expensive acquisition hardware for capture [vdPJD<sup>+</sup>14] or complex user-dependent calibrations [WBLP11, LYYB13a]. Another problem still unsolved is raised by the usage of VR headsets: prolonged partial occlusions of user's face (see Figure 1.1), making the capture of facial expressions even harder [LTO<sup>+</sup>15]. These issues of virtual characters animation and tracking devices have been pointed out by Slater [Sla14] as critical for virtual environments evolution as a mass consumer product.

Thus, this PhD thesis goal is to research and develop methods for non-expert users:

*to recognize facial movements non-intrusively  
and map them to a 3D character on-the-fly.*

## 1.2 Scientific challenges

Next, we describe the scientific challenges this PhD thesis addresses organized by thesis chapters:

- **Chapter 2: MoCap Fundamental Science:** To define which facial fea-

tures describe the face morphologies and behaviors. Due to the diversity of faces [EF75], the uniqueness of movements and the human expertise required to identify and perceive unnatural behaviors [Mor70], this knowledge emerges from the combination of deep knowledge in psychology, physiology and biomechanics. Moreover, facial features have to be selected wisely to ensure that they contain enough information to generate facial animation.

- **Chapter 3: Facial MoCap Tracking:** To recognize and track uncommon facial features, such as cheeks and forehead movements, usually not retrieved in literature approaches. The tracking is even more challenging if we want to ensure real-time performance while reducing manual intervention of the user. State of the art solutions present the following limitations: (i) the usage of intrusive markers [RE01, FL03, AM06, HCTW11]; (ii) tedious user-dependent calibrations with previous faces scans [WBLP11, LYYB13a]; (iii) expensive acquisition hardware [DHT<sup>+</sup>00, ARL<sup>+</sup>09, AFB<sup>+</sup>13, vdPJD<sup>+</sup>14, LTO<sup>+</sup>15]; (iv) generate cumulative errors when exposed to prolonged occlusions [CHZ14](Chapter 5 - MoCap VR Methods). Another challenge identified is the evaluation, testing and validation of generic CV systems. Every time researchers need to perform these tasks, they struggle to find databases that fit all system's requirements and have to define and acquire their own datasets [Bag12].
- **Chapter 4: MoCap Facial Animation:** To find the correspondence between the features tracked and the 3D character control structure, which have different space and shape configuration. This correspondence is called mapping. The mapping problem is usually solved applying example-based algorithms to the capture data or to the 3D character's control structure. Realistic real-time animations are obtained combining both example-based control structure [LWP10] and example-based capture [ARL<sup>+</sup>09, AFB<sup>+</sup>13, vdPJD<sup>+</sup>14] using complex and expensive acquisition setups [DHT<sup>+</sup>00]. The complexity and inherent costs make example-based capture and control structure approach only attainable by companies. Generally, cheapest approaches adopt only example-based captures [WBLP11, LYYB13a] reaching stable facial animations. However, in example-based capture the recognition of expressions is connected to the transfer algorithm being only able to reproduce the movements predicted by the training and contained in the blendshapes' control structure. The dependence of complex learning calibrations, lack of modularity and no reproduction of movements not predicted by the learned facial models [CHZ14], make even the alternative approaches, not suitable for non-expert users.

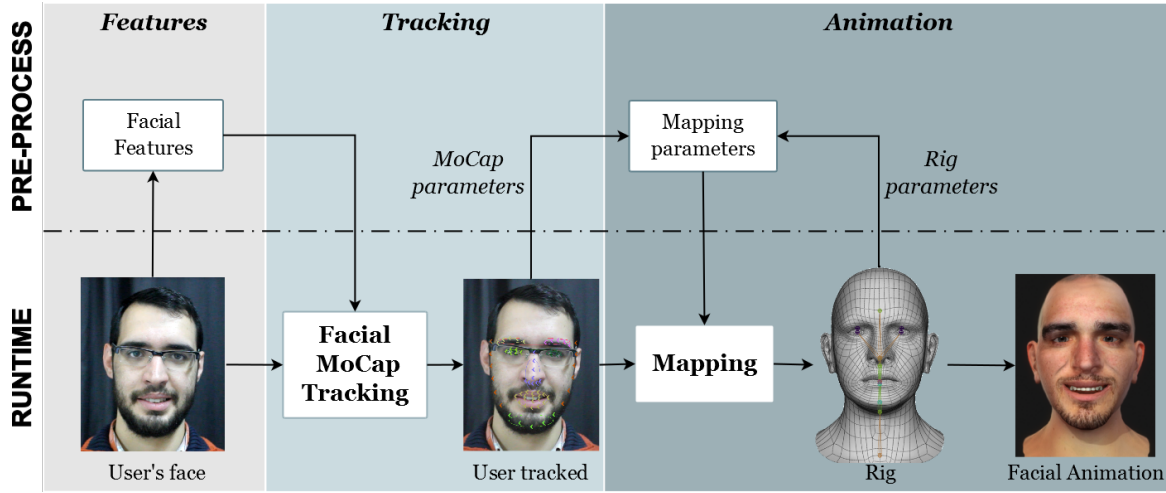


Figure 1.3: MoCap Framework: the three colored boxes represent the three research stages of this PhD thesis. The white boxes contain the methods delivered, which are distributed as pre-processes or runtime processes.

- **Chapter 5: MoCap VR Methods:** To track facial expressions when the face is partially occluded by the VR headset. Recently, Li *et al.* [LTO<sup>+</sup>15] (2015), pointed out this problem and proposed a hardware-based solution. Due to the complexity of the occlusion problem, Li *et al.*'s solution still presents limitations regarding error-accumulation in movements tracked combined with dependence of complex hardware setup using user-dependent calibration.

Technological revolution lead to the massive usage of 3D character in real-time interactive systems, like videogames, VR, augmented reality (AR), mobile applications, digital media interactions, etc. Therefore, there is an even bigger urgency in solving the aforementioned challenges to answer the demand for affordable, easy and flexible tools for 3D character creation and animation.

### 1.3 MoCap Framework

Figure 1.3 presents the MoCap framework defined in this PhD thesis. Inspired by Weise *et al.* [WBLP11] pipeline and the discoveries made in Chapter 2 of this thesis, MoCap Fundamental Science, we divide the MoCap facial animation research in three stages (Figure 1.3): facial features definition, tracking and animation. The ultimate goal is to be able to animate a facial model in real-time using off-the-shelf hardware (e.g. webcams) to capture the user. Within our pipeline it is necessary to execute two pre-processes: i) facial feature definition and ii) mapping parameters setup. In

real-time we then execute the facial feature tracking process and the mapping of the features to the 3D character.

As core stages, we have the facial MoCap tracking method (Chapter 3) and the mapping methodology for MoCap facial animation (Chapter 4). As observed in the Figure 1.3, the knowledge acquired during the MoCap Fundamental Science research (Chapter 2) was the baseline to define which features needed to be recognized (facial features at Figure 1.3) by the facial MoCap tracker (Chapter 3). The features tracked describe the unique facial morphologies and behaviors of a person which are then mapped to the 3D character using our mapping method (Chapter 4), producing believable facial animation.

The pre-processes are run one time per user to collect the following data: (i) facial features that are going to be tracked and (ii) mapping parameters with information from the topology of the user and the rig respectively defined as MoCap and rig parameters. Using the pre-processes's data and the real-time user's face image stream, the runtime processes are able to capture user's facial features using the facial MoCap Tracking algorithm and map them to a 3D character's rig (i.e. mapping) creating animation on-the-fly.

The recent technological demand from VR industry for 3D characters animation and the occlusion problem raised by VR headsets leads us to propose VR MoCap solutions (Chapter 5). Figure 1.4 shows how our VR MoCap methods fit the main MoCap framework. The MoCap VR chapter delivers the following runtime methods: (i) a generic algorithm that makes facial MoCap compatible with partial occlusions created by VR headsets; (ii) a real-time emotion recognition method and (iii) an upper face expressions prediction algorithm.

Literature methods of facial MoCap tracking and facial animation appear as unique solutions from capture to animation [LTO<sup>+</sup>15, CHZ14, vdPJD<sup>+</sup>14, LYYB13a, CWLZ13, WBLP11]. Therefore, current methods can only be applied for MoCap facial animation without persistent facial occlusions.

This PhD thesis delivers modular methods that, when combined, create a MoCap facial animation solution. But, if the methods are independent, they can integrate a variety of interactive applications, such as facial MoCap tracking for security or emotion recognition for mood-based advertising.

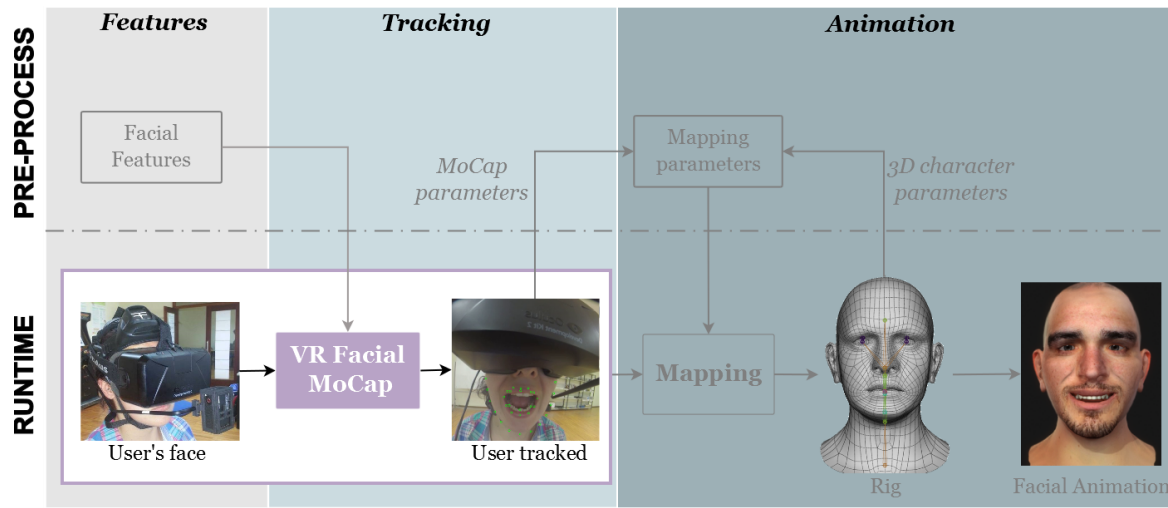


Figure 1.4: VR MoCap Framework: the purple box contains the VR MoCap methods that replace the facial MoCap Tracking in the main MoCap Framework.

## 1.4 Contributions

High quality results through facial animation with MoCap are already possible using systems like Digital Ira, from the USC-ICT [vdPJD<sup>+</sup>14]. However, they require expensive acquisition hardware [DHT<sup>+</sup>00, CHZ14].

There are alternative systems that use cheaper hardware, like Faceware<sup>1</sup>, Faceshift [LYYB13a, WBLP11] or even the recent research from Cao *et al.* [CHZ14]. After analyzing each solution, we identify the following limitations: offline manual pre-processing; scanning and calibration for user adaptation; or present incompatibilities with persistent occlusions. As an example, facial pre-scanning requirement or no support of occlusions make current approaches not suitable for real-time interactive advertising in life-like uncontrolled scenarios and VR applications for general user.

According to the scientific challenges described in 1.2, the main contributions of this PhD thesis, divided by research topic, are:

### Chapter 2 - MoCap Fundamental Science:

- a **facial features study regarding self-perception**. The experiment provides knowledge required for the definition of facial features according to our self-perception of face shape. It also highlights the importance of facial behaviors (i.e. movements) ahead of morphologies (i.e. static features that define proportions and face topology). The work was published at PLOSone journal (Appendix A).

<sup>1</sup><http://facewaretech.com/>



- a **geometric face features extraction method** inherent to facial expressions and **emotion recognition in real-time**, which resulted in a publication at VISAPP 2014 9th International Conference on Computer Vision Theory and Applications (Appendix B)

### Chapter 3 - Facial MoCap Tracking:

- a **facial MoCap method** for automatic, real-time and markerless tracking of face features not tracked by state of the art solutions [CHZ14], i.e. allows the tracking of movements in the cheeks and forehead, as an example (see VERE poster in Appendix C).
- a **database protocol**, FdMiee, to enable facial MoCap systems development and testing. FdMiee is generic and allows the simulation of a wide range of environment scenarios and facial behaviors. The protocol's validation produced a sample facial database. This work is under review in the International Journal of Computational Vision and Robotics (Appendix D).
- MoCap tracking evaluation through a **comparative study** with state of the art methods (i.e. Saragih *et al.* [SLC11a] and Cao *et al.* [CHZ14]).

### Chapter 4 - MoCap Facial Animation:

- a facial animation **mapping method** that supports **generic** MoCap systems (see VERE poster in Appendix E).

### Chapter 5 - MoCap VR Methods:

- an **algorithm that handles partial occlusions** in facial MoCap systems.
- a **real-time emotion recognition method** using only non-occluded facial features tracking (i.e. bottom part of face).
- a **predictor of facial expressions of occluded face region**, i.e. eyebrows movements, using features tracked in non-occluded region.

The three VR methods are included in the article accepted at the VISAPP 2016 11st International Conference on Computer Vision Theory and Applications (Appendix F).

The outcomes of this PhD thesis leads to the applications with the following features:

- (i) **off-the-shelf** hardware MoCap tracking
- (ii) **real-time** MoCap tracking and facial animation;
- (iii) **reduced calibration** without expertise-based manual tweaking.

The industrial contributions of our methods are evidenced by: an award of best TIC Microsoft’s award and best Business Idea in iUP25k competition (award of 15.000€); participation in Web Summit 2015 <sup>2</sup>; participation in Rockstart Summit 2015 <sup>3</sup>; an invitation from Universidade do Porto Inovação (UPIN) to participate in the Business Ignition Program for business development.

## 1.5 Application Domain

This PhD thesis gives to non-expert artists and public users the possibility of MoCap facial tracking, emotion recognition and 3D characters animation. The applicability of our methodologies are spread throughout a wide range of fields and industries, like the Entertainment, Healthcare and Education. The creation of methods compatible to VR enhance the methods’ impact in fields like marketing, learning, psychology and medical rehabilitation, where the VR technologies have been successfully applied [Sla14]. In this section, we describe the thesis potentialities in the aforementioned fields.

### 1.5.1 Entertainment Industry

Nowadays, we are familiarized with computer generated faces. They are everywhere. Besides cartoon-based productions, like *How to Train Your Dragon* (2010), *Frozen* (2013), *Big Hero 6* (2014) or *Home* (2015), the technological evolution makes possible the creation of realistic 3D characters. Examples of those characters can be found in videogames: *Assassins Creed Syndicate* (2015) and *Witcher 3: Wild Hunt* (2015), and movies: Gollum in *The Hobbit: The Battle of The Five Armies* (2014), Hulk in *Avengers: Age of Ultron* (2015) and, more recently, the character of Paul Walker’s in *Fast and Furious 7*, which was totally reconstructed after actor’s death.

The outcomes of this thesis can be applied in the context of **Broadcast** (i.e. offline applications) and **Real-time applications**:

---

<sup>2</sup><https://websummit.net/>

<sup>3</sup><http://www.rockstart.com/>



Figure 1.5: Broadcast application: Gollum face in *The Hobbit: The Battle of The Five Armies* (2014) by Weta Digital Ltd.



Figure 1.6: Real-time application: Wither's face at *The Witcher 3: Wild Hunt* (2015) by CD Project Red.

- **Broadcast:** (films, cinematic movies, trailers, short movies, promotional and advertising videos, TV commercials, etc) Due to the unlimited preprocessing time of these applications, the final 3D characters and respective animations are of high quality, relying on manual tweating and work done by expert artists. When we create facial animation using MoCap without usage of expensive acquisition hardware, like the Lightstage of Digital Ira [vdPJD<sup>+</sup>14], we are not able to reach these quality requirements. Nevertheless, our and other cheaper approaches [CHZ14, LYYB13a, CWLZ13, WBLP11] can be applied for pre-visualization, where a real-time capture and animation of characters allows the live feedback and individual shot's planning before the filming begins.
- **Real-time applications:** (videogames, VR, AR, live-shows and performances, mobile and communication software) The modular nature of our methods allows their usage for real-time character's animation in virtual environments and for emotion-based and face recognition HCI interfaces. Furthermore, our methods have special impact in VR environments. Due to their recent introduction in mass-market (Oculus VR 2014), we enable an easy user embodiment in virtual characters and on-the-fly facial animation for non-verbal communication or interaction with digital media in virtual environments.

### 1.5.2 Healthcare Sector

In Healthcare sector, our studies and methods can be applied in three different ways: face perception, expression and emotion analysis in the fields of psychology and biomechanics. Here, the fundamental science studies (Chapter 2) and emotion classification methods have the highest impact. In addition, markerless MoCap and animation method can be used in reconstruction surgery (i.e. plastic and dental) for results' pre-visualization. More recently, 3D characters animation in VR systems have been successfully applied for therapeutic purposes [PSAS13, SSV14, Sla14, GFPS11, BBRA<sup>+</sup>14]. As proof-of-concept, our methods were included in health based experiments in VERE [BBRA<sup>+</sup>14, MSSVT15] and LIFEisGAME project [AMQO13, LIF09].

### 1.5.3 Educational Interactive Systems

The educational system is currently searching for a modernization using ICT technologies. There is already proof of 3D environments' potentialities as tools for empowering learning [MSSVT15]. Thus, our methods contribute with solutions that allow

the increase of mass-scale learning interactions with embodied characters in virtual environments for socialized interactive learning. Through the dialog based systems and communication using 3D characters, we aim to dissolve geographic boundaries leading to an easy and fast way of sharing knowledge and experiences. In our journal publication "Facial emotions as a metric of 21st Century competencies: a draft protocol for acquiring and classifying facial emotion data for learning analytics" currently under review, we propose an experiment using our MoCap and emotion recognition method to access and improve learning methodologies (Appendix G).

## 1.6 Outline

The remainder of this dissertation is organized as follows:

**Chapter 2, MoCap Fundamental Science** Presents studies regarding facial features and expressions inherent to emotions. The first study includes an experiment about face features self-perception. The second study decoded which facial features are inherent to the universal emotions of Ekman and Friesen [EF75]. Then, we create a method to extract these features geometrically and use it to recognize emotions in real-time using image stream captured by a webcam as input. The emotion recognition method presents a 94% accuracy detecting the six universal emotions. This chapter provides the scientific basis for the following chapters.

**Chapter 3, Facial MoCap Tracking** Describes the novel facial MoCap method. The chapter begins with literature methods overview. Then, we propose and describe a novel method for facial tracking. Then, we present FdMiee database protocol acquisition procedure, results and conclusions taken during protocol validation. FdMiee results were further applied for facial MoCap method's evaluation and comparison with literature algorithms. Results are compared and discussed.

**Chapter 4, MoCap Facial Animation** Presents a novel mapping algorithm, which is independent of the facial MoCap tracker and, simultaneously, allows animation in real-time. Before introducing our novel mapping algorithm, a literature methods overview and respective comparison is made. Then, we describe the methodology adopted. At last, results are highlight and discussed.

**Chapter 5, VR MoCap Method** Describes a modular algorithm for MoCap that supports persistent partial occlusions. Using the occlusion support algorithm, we present two methods to access facial expressions of occluded facial regions: an emotion

recognition and an eyebrows movement predictor. Statistical validation and visual results are shown and discussed. The chapter ends with the main conclusions.

**Chapter 6, Conclusion and Future Directions** Summarizes this dissertation, highlights and discuss the work contributions, limitations and suggest ideas for future work and research questions raised.

At the end of each chapter we deliver respective results discussion and conclusions.

## Chapter 2

# MoCap Fundamental Science

*Faces. When someone approaches, the first reflex is to look into his face. The face definitively plays the main role in the so called "first impression". The face appearance, expressions and facial lines are involuntarily processed to create an idea of the person in our mind. Then, when communication starts, we stare to each other, process and use intuitively the face to decode and transmit more information about ongoing speech, mood and intentions. This give and take process is learned, trained and improved everyday by humans since early age. This human expertise makes the reproduction of appealing and believable facial animations in 3D characters a challenging task. To perfectly replicate the human face shape and movements, it requires, not only skilled artists and engineers with innovative algorithms, but also a deep knowledge regarding human perception, i.e. how we perceive human faces and facial expressions physiology and psychology. In this chapter, we define this knowledge as Fundamental Science, due to its impact in facial recognition and facial animation. We present two studies: (i) human facial perception experiment and (ii) a geometrically-based methodology for emotion recognition. Complete articles can be accessed at Appendix A and B, respectively.*

## 2.1 Face Image Task: Self-perception of Facial Features

Despite extensive research on face perception, few studies have investigated individuals knowledge about the physical features of their own face. In this study, 50 participants indicated the location of key features of their own face, relative to an anchor point corresponding to the tip of the nose, and the results are compared to the true location of the same individuals features from a standardized photograph. Horizontal and vertical errors are analyzed separately. An overall bias to underestimate vertical distances reveals a distorted face representation, with reduced face height. Factor analysis is used to identify separable subconfigurations of facial features with correlated localization errors. Independent representations of upper and lower facial features emerge from the data pattern. The major source of variation across individuals is in the representation of face shape, with a spectrum from tall/thin to short/wide representation. Visual identification of ones own face is excellent, and facial features are routinely used for establishing personal identity. However, our results show that spatial knowledge of ones own face is remarkably poor, suggesting that face representation may not contribute strongly to self-awareness.

These conclusions make us understand that humans do not have a spacial knowledge of his own static face features distribution. However, it remains open an analysis of moving facial features as future work. The discoveries achieved allowed us to understand the minimum facial features (Figure 2.1 A) for one's face representation. So, the perception experiment contributed with the knowledge required for the definition of which facial features are going to be tracked by the facial MoCap method (Chapter 3) with direct impact in the animation delivered by the mapping method (Chapter 4). Moreover, it also decodes that we have different representations of upper and bottom parts of the face. The last statement is useful to understand the influence of face partial occlusions in face shape's perception. The occlusions' topic is addressed at Chapter 5 - MoCap VR methods.

**Contribution:** This study was executed in partnership with psychology researchers from University College of London (UCL), United Kingdom. We were responsible for the experiment technology definition, analysis and results generation, including face deformations and visual representations.



### 2.1.1 Background

Face perception is a central topic in modern psychology. The field has overwhelmingly used visual stimuli and focused on face recognition, even when considering perception of ones own face [UKMS<sup>+</sup>05]. People see their own face only rarelyvanishingly rarely until the recent ready availability of mirrors. Nevertheless, several studies indicate a specific mechanism involved in recognizing ones own face (e.g., [RKB12], see also [DB11] for a review). Much of this literature focus on sensitivity to facial symmetry and its relation to mirror effects [Bré03, BCF05], and cerebral hemispheric specialization [BCF04]. Many visual face recognition studies suggest a superior and accurate visual representation of ones own face [DB11]. However, the persistence of this advantage even when faces are inverted suggests that it relies on local rather than configural processing [KB10]. In general, the self-face visual recognition literature cannot readily distinguish between self-face processing based on familiarity with a visual image of ones own face suitable for template matching, or based on structural knowledge about what ones face is like i.e., a face image or a hypothetical stored representation containing information about the positions of facial features relative to one another, akin to the body structural description [CDHRF08]. Here we largely remove the visual recognition aspect of self-face processing to focus on the latter, structural representation aspect. Only one study has investigated somatosensory self-face perception [FPL<sup>+</sup>13], and found generally poor performance. Therefore, it remains unclear what people know about their own facial structure, and how this knowledge is stored and represented independent of a specific visual stimulus.

Recently, there were developed tasks for studying the sensed position of body parts [LH10], and stored models of ones own body [FPL<sup>+</sup>13, LH12]. These representations both showed systematic patterns of distortion, which potentially indicate how spatial information about bodies is represented and stored in the brain. Here we report results on representation of ones own facial features using a method that does involve visual recognition. We show, first, that people make large errors in locating their own facial features, particularly underestimating face height. Second, we show through factor analysis that the representation of facial feature locations follows a characteristic structure. The patterns of localization errors showed covariance across specific subsets of features, which may be relevant to identifying the organization of face representation at a supra-featural, or configural level. The overall structure of face representations implies an important distortion of face shape. Our work provides a novel and systematic approach to a classic question of Gestalt psychology: how are configurations of multiple features represented in the brain as a composite pattern?

Our results may also be relevant to the considerable concern regarding ones own facial structure and appearance in some individuals and cultures.

### 2.1.2 Methodology

**Ethics Statement** All participants gave informed written consent. All experiments were approved by the local ethics committee at UCL. Participants were seated in front of a computer screen in portrait orientation (Dell model 2007 WFPb, measuring 43.5 cm vertical, 27.5 cm horizontal) which displayed only a small central dot. The position of the dot on the screen was randomized across trials. Participants were instructed to imagine their own face projected frontally, life-size on the screen, with the tip of the nose located at the dot. They used a mouse to indicate the locations corresponding to 11 landmark facial features. The figure reproduced as Figure 2.1 A was shown to participants before the experiment to indicate the exact anatomical landmarks intended. Before each trial, a text label (e.g., bottom of chin, center of left eye) briefly appeared centrally on the screen. Environmental lighting was controlled so that they could not see any reflection of their face on the screen. Each landmark was judged five times in a random order. To quantify errors in perceived position of facial features, responses were later compared to the actual locations of those landmarks, obtained by taking a photograph under standardized conditions and rendering it at life-size on the same screen. The average horizontal (x) and vertical (y) error for attempts to locate each facial landmark were calculated.

In this experiment took part 50 participants (24 female, average age 25 years). The X-axis data from left-sided landmarks (ears, nose and mouth edges, eyes) was reflected in the midline, and averaged with the corresponding right-sided landmark. This imposed an assumption of facial symmetry, but reduced the number of dependent variables and avoided possible confusion regarding the terms left and right in the context of the task. By analyzing the pattern of errors, we aimed to investigate the internal stored representation of ones own face. Finally, a subset of 10 participants were asked to attend for a second session, in which the screen was rotated to landscape mode.

### 2.1.3 Results

The average error vectors are shown superimposed on a schematic face in Figure 2.1. They reveal large overall biases in locating facial landmarks. The anatomical structure of the face is very different in the horizontal and vertical dimensions. The horizontal

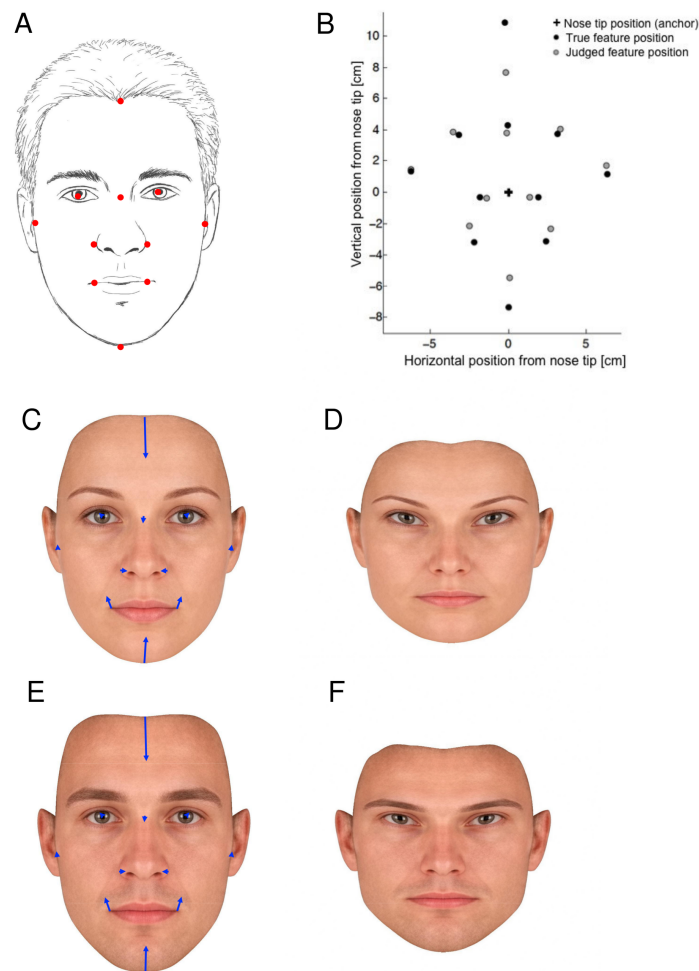


Figure 2.1: Biases in face representation: A: Schematic of feature locations used to instruct participants. B: Actual and mean represented locations. C: Average of 50 female faces reproduced with permission from PerceptionLab [Per]. Blue arrows indicate mean judgment error for each feature. D: Average female face adjusted according to the mean represented locations of our participants. E, F: as for C, D with average of 50 male faces.

dimension is characterized by symmetry and homology, while the vertical dimension lacks both these attributes. Therefore, we expected different patterns of error in the X and Y dimensions, and accordingly analyses each dimension separately. In the horizontal dimension, mouth and eye width are overestimated, while nose width is underestimated. In the vertical dimension, the hairline is represented as lower, and the chin as higher, than their true locations, suggesting that the face is represented as shorter than its true height. No simple geometric distortion can explain the overall pattern of biases: for example, the compression of face height may appear to be a regression of judgment towards the mean defined by the anchor point on the nose tip. However, eye and ear vertical positions appear to be unaffected by this bias, and the bias is absent in the horizontal dimension, suggesting it is not simply a matter of eccentricity. Moreover, Bonferroni-corrected testing showed significant biases for some facial features close to the anchor point, but not for those farther away.

Part	Mean Horizontal Error cm(SD)	Mean Vertical Error cm(SD)
<b>Hairline</b>	-0.0875 (0.2989)	<b>-3.1533 (1.8734)</b>
<b>Chin</b>	-0.0640 (0.3028)	<b>1.8987 (1.6650)</b>
<b>Ear</b>	0.0396 (1.5981)	0.3534 (1.7092)
<b>Nose Bridge</b>	<b>0.0735 (0.1401)</b>	-0.4734 (1.3835)
<b>Nose</b>	<b>0.4995 (0.6141)</b>	-0.0246 (0.5963)
<b>Mouth</b>	-0.3170 (1.0228)	<b>0.9060 (0.9188)</b>
<b>Eyes</b>	-0.2510 (1.0588)	0.2509 (1.4582)

Table 2.1: Average localization errors for each feature in cm.

In the 10 participants who performed the task with the screen in portrait and landscape mode, we found no effects of screen orientation on judgement error, and no interaction between screen orientation and feature judged, in either X or Y dimensions (all  $F < 1$ , all  $p > 0.60$ ). To investigate the underlying structure of the face representation shown in Figure 2.1, we applied separate factor analyses to x and y judgment errors (see Appendix A - Section: Supporting information). The ratio of measurements-to-cases falls within the guideline range for exploratory factor analysis [MSK05]. Principal components were extracted, and *varimax* rotated. Factors with eigenvalues over 1 were retained (Table 2.2 and Figure S1 in the Appendix A - Section: Supporting information).

For horizontal errors, we identified three retainable factors, which we label X1, X2, X3 for convenience, corresponding to the principal, independent sources of variability in horizontal judgment errors for facial features. The first factor (X1) suggested

Factor	X1	X2	X3	Y1	Y2
<b>Eigenvalue</b>	2.75	1.70	1.05	3.09	1.84
<b>Variance proportion</b>	39%	24%	15%	44%	26%
<b>Hairline</b>	-0.00134	0.91808	-0.08091	0.86393	-0.14580
<b>Chin</b>	-0.02570	-0.01205	0.97501	-0.45231	0.76155
<b>Nose bridge</b>	0.09507	0.89391	0.07555	0.89044	-0.14270
<b>Nose edge</b>	0.66612	-0.24710	-0.11494	0.21907	0.77971
<b>Mouth</b>	0.88904	0.09058	-0.14932	-0.17919	0.92808
<b>Eye</b>	0.90951	0.14324	0.01498	0.92128	-0.02805
<b>Ear</b>	0.78575	0.14074	0.23051	0.32930	0.24252

Table 2.2: Factor scores for horizontal X and vertical Y components.

a tendency to expand facial width outward from the midline. It loaded strongly and roughly equally on all lateralized structures (eye, mouth, ear, nose), but not on midline structures (center of hairline, bridge of nose, chin). The second factor (X2) suggested lateral distortion of the upper face. It loaded largely on the hairline and nose bridge. The third factor (X3) suggested lateral distortion of the lower face, loading almost exclusively on the chin. For analysis of vertical errors, only two factors were retained. The first (Y1) loaded strongly on upper face structures (eyes), including midline structures (nose bridge, hairline), but with some modest negative loading on the chin. This factor suggested a vertical expansion of the face from its center. The loadings of the second factor (Y2) on lower face structures (mouth, nose edges, chin) suggest a vertical shift confined to the lower face.

These factor solutions carry important information about the internal structure of horizontal and vertical face representation. Factors X1, X2, Y1 and Y2 all loaded on more than one facial feature. The loading patterns suggest complexes of two or more individual features that group together, and which covary across the face representations of different individuals. By this means, we could identify separable representations of lateral and midline horizontal facial features, and separable representations of upper and lower face vertical structure. The effects of varying each factor on an average face are shown as vectors in Figure S1 and pictorially in Figure S2 both in the Appendix A.

We also investigated the overall geometry of face representation by seeking an inter-domain association between factors affecting horizontal and vertical errors. We used canonical correlation to identify the principal associations between our horizontal factors (X1, X2) and vertical factors (Y1, Y2).

The first canonical variate accounted for 48.5% of the variance between the horizontal

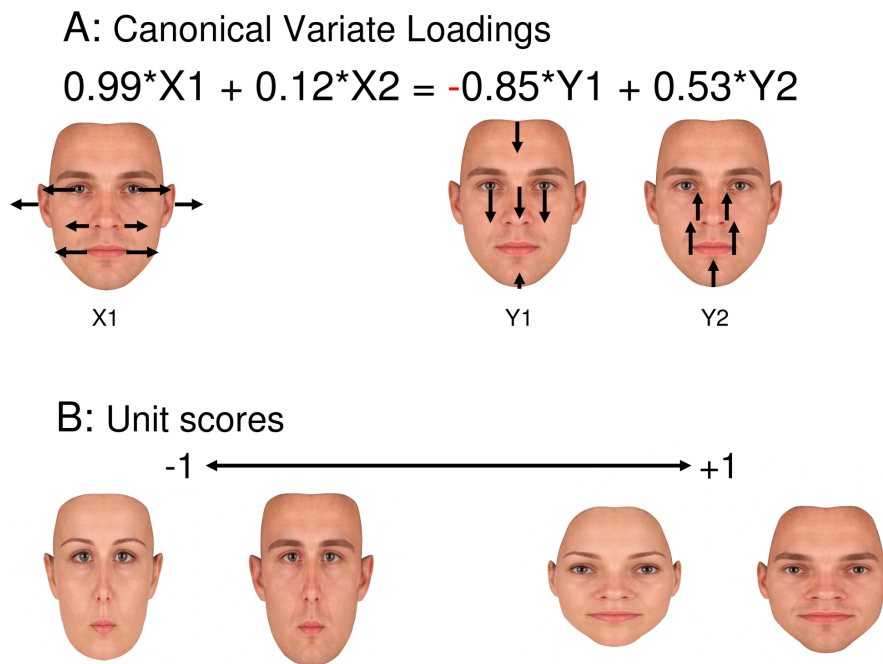


Figure 2.2: Association between horizontal and vertical distortion factors demonstrates variation in representation of face shape across individuals. Results of a canonical correlation between the horizontal (X1,X2) and vertical (Y1,Y2) factors. A: Vectors showing the principal feature loadings ( $>0.4$  or  $<-0.4$ ) of the factors, adjusted by the coefficients indicating important ( $>0.4$  or  $<-0.4$ ) contributions to the canonical variate. The vector lengths are shown at 4x the actual values for visual clarity. Note the negative sign for Y1 coefficient. B: Average female and male faces implied by a low and high score on the canonical variate. Note that the canonical variate separates long and thin from short and wide face representations.

and vertical factors and was highly significant (Wilks Lambda 0.506, approximated by  $F(4,92)=9.34$ ,  $p<0.001$ ). The standardized weights showed that the canonical variate related X1 (weighting 0.99) negatively to Y1 (-0.85) and positively, though less strongly, to Y2 (0.53). In contrast, factor X2 made little contribution to this inter-domain association (weighting 0.12), suggesting that it constituted an independent aspect of facial structure. The combination of weightings in the first canonical variate is readily interpretable as face aspect ratio, or 2D shape. The lateral shift of eyes, mouth edges, ears and nose captured by factor X1 was associated with a downward shift of the hairline and nose-bridge (captured by Y1), and some upward shift of the mouth, nose edges and chin (captured by Y2). That is, the lateral expansion of the face was strongly associated with a vertical compression of towards the face center, suggesting that the face aspect ratio is the major structural principle of face representation. The second canonical variate explained only 1.7% of the shared variance between factors, and was far from significant ( $p=0.37$ ). Factor X3 was excluded from the inter-domain analysis, as its loading was largely confined to a single feature. However, re-running the analysis with this factor included had only small effects on weightings of inter-domain association and did not change the pattern of inference. Figure 2.2 A shows the vectors associated with the major loadings ( $>0.4$ ) of each factor, adjusted by the factors weighting in the canonical variate. Figure 2.2 B shows the face images implied by a positive and negative unit score on the canonical variate.

#### 2.1.4 Discussion and Conclusions

For a more complete analysis and overview (i.e. in a psychological point of view), we suggest the full reading of the article "Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation" at Appendix A.

Distortions in face representation have been widely reported in visual perception. For example, one study using adaptation procedures investigated aspect suggested that aspect ratio was a core component of face coding in the human brain [WC03]. However, those studies did not specifically test for other distortions of face coding, apart from shape, and could design only a limited range of stimuli to test dimensions of coding hypothesized a priori. In our approach, by contrast, the key dimensions of face coding emerge from the pattern of participants responses, rather than by experimenters choice of stimulus set.

This study proposes a new method to access the stored knowledge about the self image of our face and structural arrangement of one's own facial features. Importantly,

this method allows the structural description of the face independently of visual recognition. Analyzing results, the first prominent feature was the aspect ratio defined by facial features. Our data therefore provides strong and independent convergent evidence that aspect ratio is a major source of variation in face representation. Not only are people poor at estimating the shape of their own face (Figure 2.1) but the principal source of variation across individuals is in the biased representation of face shape. The last conclusions and knowledge acquired during the experiment influenced the definition of the facial features tracked by the facial MoCap tracking method at Chapter 3, which have impact in the quality of the animation generated at Chapter 4.

A second clear statement is related to the differences between upper and lower facial features. For the majority of the factors extracted, we observe high loadings on the upper face accompanied by low loadings of the lower face, or vice versa. We explored this discovery during the deployment of the VR MoCap methods (Chapter 5). We made a research of the influence of partial occlusions of the face, specially upper face occlusion, and found that previous conclusion is confirmed by several recent emotional studies [EA11, BSSM<sup>+</sup>13]. To a detailed study of occlusions' topic, we forward the reader to Chapter 5.

The take home message is: human's spatial knowledge of his own face is remarkably poor. Though, our analysis was executed exclusively using static facial features, which leaves as future work the research of moving facial features and their impact in face self-perception.



## 2.2 Real-Time Emotion Recognition

This section presents a study to decode and define which facial features are inherent to the six universal emotions [EF75] (Joy, Sorrow, Surprise, Fear, Disgust and Anger) plus Neutral and Contemptuous. Using these features, we deployed a feature extraction method and a real-time emotion classifier. Compared to literature, the methodology outperformed accuracy (i.e. 94%) and, simultaneously, solves four of the facial emotion recognition issues [Bet12]. In this section, we selected directly the parts of original work that have direct impact in this PhD thesis: background, features definition and extraction methodology, main results and respective discussion and conclusions. The complete study can be accessed at Appendix B.

**Contribution:** This second study was performed in partnership with PERCRO Laboratory from Scuola Superiore Sant’Ana, Pisa, Italy. We were responsible for the feature definition, geometrical extraction method deployment and machine learning validation procedures.

### 2.2.1 Background

Facial expressions play a crucial role in communication and interaction between humans. In the absence of other information such as speech interaction, facial expressions can transmit emotions, opinions and clues regarding cognitive states [KS10]. A fully automatic real-time face features extraction for emotion recognition allows to enhance the communication realism between humans and machines. Several face recognition systems have been developed for real time facial features detection as well as (e.g. [BLFM03]). Psychological studies have been conducted to decode this information only using facial expressions, such as the Facial Action Coding System (FACS) developed by Ekman [EF78]. As stated on the recent survey [JN12], among existing facial expression recognition systems, the common three-step pipeline for facial expressions classification [Bet09] is composed by:

1. the *Facial recognition* stage;
2. the *Features extraction* stage;
3. the *Machine learning classifier* stage (preliminary model training and online prediction of facial emotions).

As claimed in the same survey, the second pipeline stage (features extraction) strongly affects the accuracy and computational cost of the overall system. It follows that the features to be extracted and corresponding extraction methods are fundamental for the overall performances. The commonly used methods for feature extraction can be divided into geometrical methods (i.e. features are extracted from shape or salient point locations such as the mouth or the eyes [KQP03]) and appearance-based methods (i.e. skin features like frowns or wrinkles, Gabor Wavelets [Fis04]).

Geometric features are selected from landmarks positions of essential parts of the face (i.e. eyes, eyebrows and mouth) obtained by technique of face features recognition. These extraction methods are characterized by their simplicity and low computational cost, but their accuracy is extremely dependent on the face recognition performances. Examples of emotion classification methodologies that use geometric features extraction are [CK09, NAHF<sup>+</sup>12, GXhJlXg09, Ham07, SAK04, KP07]. However, high accuracies on emotion detection usually require a calibration with a neutral face ([KP07, GXhJlXg09, NAHF<sup>+</sup>12, CK09, Ham07]), an increase of the computational cost ([GXhJlXg09, SAK04]), a decrease of the number of emotions detected ([NAHF<sup>+</sup>12, Ham07]) or a manual grid nodes positioning [KP07]. On the other hand, appearance-based features work directly on image and not on single extracted points (e.g. Gabor Wavelets [KBP08] and Local Binary Patterns [SGM09] [CS10]). They usually analyze the skin texture, extracting relevant features for emotion detection. Involving a long learning process for facial model generation, the appearance feature method becomes more complex than the geometric approach, compromising also the real-time feature required by the process (appearance-based features show high variability in performance time from 9.6 to 11.99 seconds [ZTC12]).

Hybrid approaches, that combine geometric and appearance extraction can be found (i.e. [YA11]) with higher accuracies, but they are still characterized by a high computational cost. The aim of this research work is to propose a feature extraction method that provides performances comparable with appearance-based methods without compromising the real-time and automation requirements of the system. Nevertheless, we intent to solve the following main four facial emotion recognition issues [Bet09]:

1. real-time requirement: communication between humans is a real time process with a time scale order of about 40 milliseconds [BLFM03];
2. capability of recognition of multiple standard emotions on people with different anthropometric facial traits;
3. capability of recognition of the facial emotions without neutral face comparison

calibration;

4. automatic self-calibration capability without manual intervention.

(equivalent optimizations of these four issues can also be extracted from *Jamshidnezhad et al.*'s survey [JN12]). Real-time issue is solved using a low complexity features extraction method without compromising the accuracy of emotion detection. In order to show the capacity of the second issue, we test our system on a multi-cultural database, the Radboud face database [LDBW], featured with multiple emotions traits [Bet09]. Additionally, we investigate all six universal emotions[EF75] (Joy, Sorrow, Surprise, Fear, Disgust and Anger) plus Neutral and Contemptuous. Regarding the third issue, though with slightly lower performance relative to neutral face comparison calibration, our method allows the recognition of eight different emotions without requiring any calibration process. To avoid any manual intervention in the localization of the seed landmarks required by our proposed geometrical features, we use as reference example in this work, a marker-less facial landmark recognition and localization software based on the Saragih's MoCap approach [SLC11b]. However, face recognition can be done with the use of different marker-based and marker-less systems which allow the localization of the basic landmarks defined in our system for emotion classification. Therefore, as main contribution, we defined facial features inherent to emotions and proposed a method for their extraction in real time, for further emotion recognition.

### 2.2.2 Methodology:

#### Geometric Features Extraction Method

In this work, we propose a set of facial features suitable for marker-based and marker-less systems, that included, as basis, the 11 facial landmarks defined in previous study.

In fact, we present an approach to extract facial features that are truly connected to facial expression. We start from a subset composed by 19 elements (see Fig. 2.3 and Table 2.3) of the 54 anthropometric facial landmarks set defined in [LBJ11] that are usually localized using facial recognition methods.

By this date (2013), the testing benchmark used for our extraction method is an existing marker-less system for landmark identification and localization by Saragih *et al.* [SLC11a]. Their approach reduces detection ambiguities, presents low online computational complexity and high detection efficiency outperforming the other popular deformable real-time models to track and model non-rigid objects (Active Appearance

Table 2.3: The subset of anthropometric facial landmarks used to calculate our proposed geometric facial features.

No.	Landmark	Label	Region
1	Right Cheilion	$A_M$	Mouth
2	Left Cheilion	$B_M$	Mouth
3	Labiale Superius	$U_{m1}$	Mouth
4	Labiale Inferius	$D_{m2}$	Mouth
5	Left Exocanthion	$Ell_M$	Left Eye
6	Right Exocanthion	$Elr_M$	Left Eye
7	Palpebrale Superius	$UEL_{m3}$	Left Eye
8	Palpebrale Inferius	$DEl_{m4}$	Left Eye
9	Left Exocanthion	$Erl_M$	Right Eye
10	Right Exocanthion	$Err_M$	Right Eye
11	Palpebrale Superius	$UEr_{m5}$	Right Eye
12	Palpebrale Inferius	$DEr_{m6}$	Right Eye
13	Zygofrontale	$EBll_M$	Left Eyebrow
14	Inner Eyebrown	$EBlr_M$	Left Eyebrow
15	Superciliare	$UEBl_{m7}$	Left Eyebrow
16	Inner Eyebrown	$EBrl_M$	Right Eyebrow
17	Zygofrontale	$EBrr_M$	Right Eyebrow
18	Superciliare	$UEBr_{m8}$	Right Eyebrow
19	Subnasale	$SN$	Nose

Models (AAM) [ASWG09], Active Shape Models (ASM) [CC92], 3D morphable models [BV99a] and Constrained Local Models (CLMs) [CC]). For a more complete and recent state of the art tracking methods, we forward the reader to Background section of the Chapter 3 - facial MoCap.

Saragih *et al.* [SLC11a] system identifies and localizes 66 2D landmarks on the face. Through the repetitive observation of facial behaviors during emotion expressions, we empirically choose a subset of 19 facial landmarks that better capture these facial changes among the 66 facetracker ones. Note that these features are an extension from the 11 minimum facial features use to describe face shape, since we are focusing in motion description instead of static shape.

Using the landmark positions in the image space, we define two classes of features: *eccentricity* and *linear* features. These features are normalized to the range [0,1] to let the feature not affected by people anthropometric traits dependencies. So, we extract geometric relations among landmark positions during emotional expression for people with different ethnicities and ages.

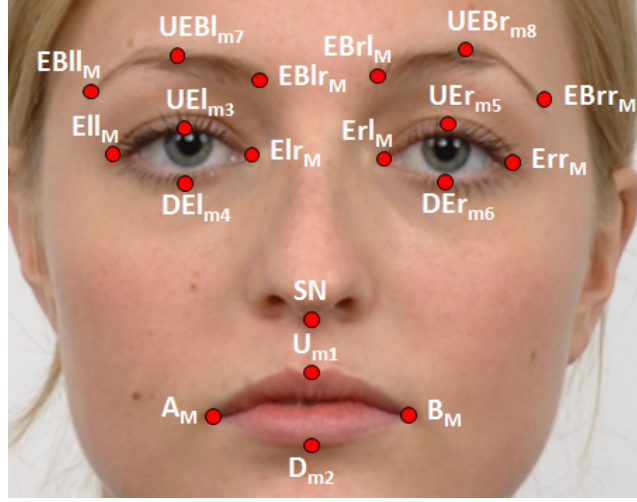


Figure 2.3: The subset composed by 19 points of the 66 facial landmarks used to extract our proposed geometric facial features.

### 2.2.2.1 Eccentricity Features

The *eccentricity* features are determined by calculating the eccentricity of ellipses constructed using specific facial landmarks. Geometrically, the *eccentricity* measures how the ellipse deviates from being circular. For ellipses the *eccentricity* is higher than zero and lower than one, being zero if it is a circle. As example, drawing an ellipse using the landmarks of the mouth, it is possible to see that while smiling the eccentricity is higher than zero, but when expressing surprise it is closer to a circle and almost zero. A similar phenomenon can be observed also in the eyebrow and eye areas. Therefore, we use the eccentricity to extract new features information and classify facial emotions. More in detail, the selected landmarks for this kind of features are 18 over 19 (see Table 2.3 and Figure 2.3), whereas the total defined *eccentricity* features are eight: two in the mouth region, four in the eye region and two in the eyebrows region (more details can be found in Table 2.4). Now, we describe the *eccentricity* extraction algorithm applied to the mouth region. The same algorithm can be simply applied to the other face areas (eyebrows and eyes) following the same guidelines.

With reference to Figure 2.4.a, let  $A_M$  and  $B_M$  be the end points of the major axis corresponding to the side ends of the mouth, while  $U_{m1}$  the upper end points of the minor axis (the distance between the major axis and  $U_{m1}$  corresponds to the semi-minor axis). Of course, the symmetry of  $U_{m1}$  with respect to  $A_M$  and  $B_M$  is not assured. For this reason, in the following, we will refer to each ellipse as the *best fitting ellipse* among the three points having the semi-minor axis equal to the distance between  $U_{m1}$  and the line  $A_MB_M$ .

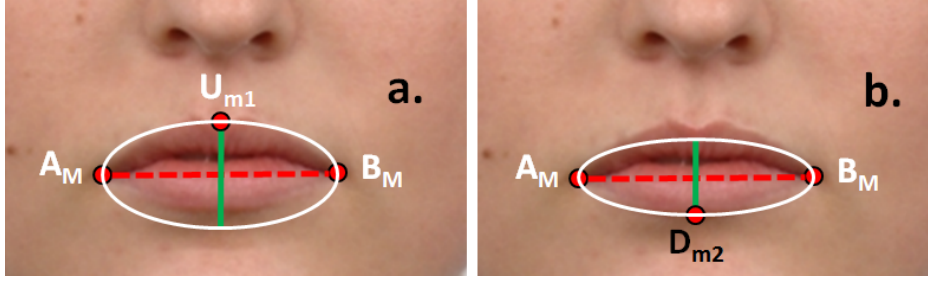


Figure 2.4: The definition of the first (a.), "upper" and the second (b.), "lower" ellipses of the mouth region using respectively the triple  $(A_M, B_M, U_{m1})$  and  $(A_M, B_M, D_{m2})$ .

We construct the first ellipse  $E_1$ , named "upper" ellipse, defined by the triple  $(A_M, B_M, U_{m1})$  and calculate its eccentricity  $e_1$ . The eccentricity of an ellipse is defined as the ratio of the distance between the two foci, to the length of the major axis or equivalently:

$$e = \frac{\sqrt{a^2 - b^2}}{a} \quad (2.1)$$

where  $a = \frac{B_{Mx} - A_{Mx}}{2}$  and  $b = A_{My} - U_{m1y}$  are respectively one-half of the ellipse  $E$ 's major and minor axes, whereas  $x$  and  $y$  indicate the horizontal and the vertical components of the point in the image space. As mentioned above, for an ellipse, the eccentricity is in the range  $[0,1]$ . When the eccentricity is 0, the foci coincide with the center point and the figure is a circle. As the eccentricity tends toward 1, the ellipse gets a more elongated shape. It tends towards a line segment if the two foci remain a finite distance apart and a parabola if one focus is kept fixed as the other is allowed to move arbitrarily far away.

We repeat the same procedure for the ellipse  $E_2$ , named "lower" ellipse, using the lower end of the mouth (see Figure 2.4.b). The other six ellipses are, then, constructed following the same extraction algorithm using the features summarized in Table 2.4 (for the landmark labels refer to Table 2.3 and Figure 2.3). It is clear that for both eyebrows, it is not possible to calculate the lower ellipses due to their morphology. The final results of the ellipse construction can be seen in Figure 2.5.a, whereas in Figure 2.5.b it is possible to see how the eccentricities of the facial ellipses changes according to the person's facial emotion.

### 2.2.2.2 Linear features

The *linear* features are determined by calculating linear distances between couples of landmarks normalized with respect to a physiologically greater facial inter-landmark

Table 2.4: The eight ellipses used to extract the eccentricity features (for the landmark labels please refer to Figure 2.3).

Ellipse	Point Triple	Region
$E_1$	$(A_M, B_M, U_{m1})$	Upper mouth
$E_2$	$(A_M, B_M, D_{m2})$	Lower mouth
$E_3$	$(Ell_M, Elr_M, UEl_{m3})$	Upper left eye
$E_4$	$(Ell_M, Elr_M, DEL_{m4})$	Lower left eye
$E_5$	$(Erl_M, Err_M, UEr_{m5})$	Upper right eye
$E_6$	$(Erl_M, Err_M, DEr_{m6})$	Lower right eye
$E_7$	$(EBll_M, EBlr_M, UEBl_{m7})$	Left eyebrow
$E_8$	$(EBrl_M, EBrr_M, UEBr_{m8})$	Right eyebrow

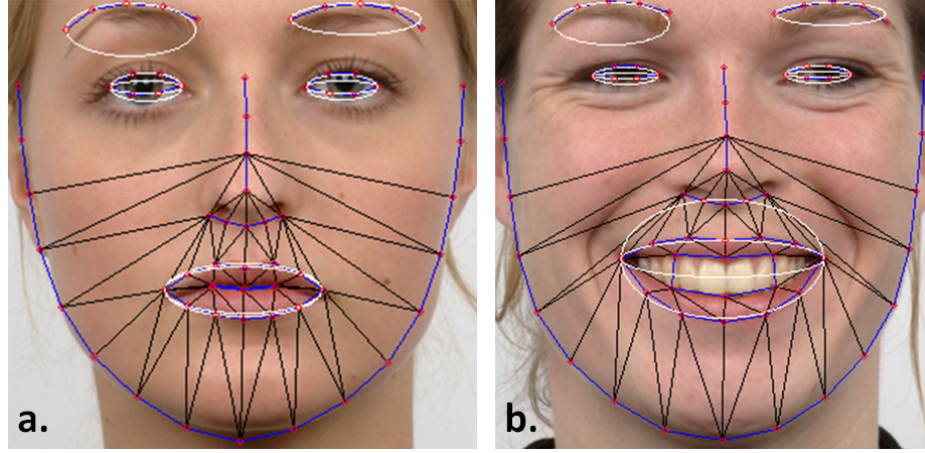


Figure 2.5: The final results of the eight ellipse construction (a). Eccentricities of the facial ellipses changes according to the person's facial emotion (b).

distance. These distances intend to quantitatively evaluate the relative movements between facial landmarks while expressing emotions. The selected distances are those corresponding to the movements between eyes and eyebrows  $L_1$ , mouth and nose  $L_2$  and upper and lower mouth points  $L_3$ . More in detail, with reference to Table 2.3 and Figure 2.3, indicating with  $_y$  only the vertical component of each point in the image space and selecting as  $DEN = \overline{UEl_{m3y}SN_y}$  the normalizing distance, we calculate a total of three linear features as:

$$1. L_1 = \overline{UEBl_{m7y}UEl_{m3y}}/DEN;$$

$$2. L_2 = \overline{U_{m1y}SN_y}/DEN;$$

$$3. L_3 = \overline{D_{m2y}SN_y}/DEN;$$

### 2.2.2.3 Extracted Features

In order to fully evaluate and compare our defined features, we consider five types of feature subsets:

1. only linear features (subset  $S1$ : 3 elements);
2. only eccentricity features (subset  $S2$ : 8 elements);
3. both eccentricity and linear features (subset  $S3$ : 11 elements);
4. differential eccentricity and linear features with respect to those calculated for Neutral emotion face (subset  $S4$ : 11 elements);
5. all features corresponding to the union of  $S3$  and  $S4$  (subset  $S5$ : 22 elements).

(where the differential features are calculated as:

$$df_{i,x} = f_{i,x} - f_{i,neutral}$$

with  $i$  representing a subject of the database and  $x$  an emotion), resulting in a total number of calculated features for the entire database equal to  $(1.385 \text{ pictures} \times 22 \text{ } S5 \text{ numerosity}) 30.470$ . The five subsets can be grouped into two main classes:

1. the *intra-person-independent* or *non-differential* subsets  $S1$ ,  $S2$  and  $S3$  that do not require any kind of calibration with other facial emotion states of the same person;
2. the *intra-person-dependent* or *differential* subsets  $S4$  and  $S5$  that require a calibration stage using the Neutral emotion of the same person.

### 2.2.3 Classifier validation

The *classifier validation* test is subdivided into two parts:

1. the *training phase* of three emotion classification methods (k-Nearest Neighbours, Support Vector Machine and Random Forests that will be described in detail later);



2. the *classifier accuracy estimation* of the three methods in order to identify the best classification method to be used in the *second experiment*.

Both for training and for the accuracy estimation, we used only the subset  $S5$ , that is the most inclusive feature subset. In order to train a classifier according to supervising learning approach, we need an input dataset containing rows of features and an output class (e.g. the emotion). The trained classifier provides a model that can be used to predict the emotion corresponding to a set of features, even if the classifier did not use these combinations of features in the training process. According to [ZPRH09], the most significant classifiers that can be used for our experiment are *k-Nearest Neighbours* [CH67], *Support Vector Machine* [AW99] and *Random Forests* [Bre01]. As mentioned above, as final result of the classifier validation, we will select the classification method that provides best performances on emotion recognition accuracy using only the subset  $S5$ .

Regarding the second part of the first test, to quantify the classification accuracy of the three presented methods, we use the K-Fold Cross Validation Method (K-Fold CRM). More in detail, the k-Fold CRM, after having iterated  $k$  times the process of dividing a database in  $k$  slices, trains a classifier with  $k - 1$  slices. The remaining slices are used as test sets on their respective  $k - 1$  trained classifier to calculate the accuracy and provides as final accuracy value the average of the  $k$  calculated accuracies.

In our case, we impose  $K = 10$ , because this is the number that provides statistical significance to the conducted analysis [RPL10]. The accuracy estimations obtained with the three investigated methods, k-Nearest Neighbours (with  $k = 1$ ), Support Vector Machine and Random Forests using the subset  $S5$  to recognize all eight emotions are the following, 85%, 88% and 89%, respectively.

Due to its better performances, we decided to use only the Random forests classifier to conduct the second experiment, that is a full analysis considering all the feature subsets and four different subsets of emotions with numerosity equal to 6, 7 and 8 emotions.

### 2.2.3.1 Features Evaluation

The results of the full analysis conducted using the Random Forests classifier (selected after the classifier validation test) are reported in Table 2.6. As expected,  $S4$  and  $S5$  provided better recognition performances with an overall accuracy increment of 6% (in the 6 emotions test) and of 9% (in the 8 emotions test) with respect to that obtained

No. tested emotions	$S1[\%]$	$S2[\%]$	$S3[\%]$	$S4[\%]$	$S5[\%]$
8	51	76	80	86	89
7*	61	80	84	88	<b>90</b>
7**	60	81	84	<b>90</b>	<b>92</b>
6***	67	87	89	<b>91</b>	<b>94</b>

Figure 2.6: Results using a Random Forests classifier for each dataset composed by a sub-set of features of a subset of emotions to classify. \* means without considering contemptuous emotion, \*\* without considering neutral emotion, \*\*\* without considering neutral and contemptuous emotions

Method	Differential	Accuracy[ % ]
<i>Michel et al. (Michel and El Kaliouby, 2003)</i>	No	72
<i>Pardàs et al. (Pardàs and Bonafonte, 2002)</i>	No	84
<i>Bartlett et al. (Bartlett et al., 2003)</i>	No	84
<b>Our method <math>S3</math></b>	No	<b>89</b>
<i>Michel et al. (Michel and El Kaliouby, 2003)</i>	Yes	84
<i>Cohen et al. (Cohen et al., 2003)</i>	Yes	88
<i>Wang et al. (Wang and Yin, 2007)</i>	Yes	93
<b>Our method <math>S5</math></b>	Yes	<b>94</b>

Figure 2.7: Accuracy comparison of emotion facial recognition methods(not differential or differential features) with six universal emotions.

	Angry	Cont.	Disgust	Fear	Joy	Neutral	Sorrow	Surprise
<b>Angry</b>	<b>76   84   89</b>	09   06   03	02   03   02	00   00   00	00   00   00	07   01   00	05   06   06	00   00   00
<b>Cont.</b>	06   01   02	<b>73   77   82</b>	01   01   01	01   00   00	04   01   00	08   11   08	08   09   09	00   00   00
<b>Disgust</b>	05   03   01	01   00   00	<b>91   89   94</b>	00   01   01	02   04   00	01   01   01	01   02   05	00   01   00
<b>Fear</b>	01   00   00	00   02   01	00   00   00	<b>82   87   87</b>	00   00   00	05   02   01	05   04   05	08   07   07
<b>Joy</b>	02   00   00	01   01   00	01   04   02	00   00   00	<b>95   94   97</b>	01   00   00	01   02   00	00   00   00
<b>Neutral</b>	03   00   00	11   08   06	02   00   00	06   03   01	00   00   00	<b>69   84   87</b>	08   05   03	00   00   00
<b>Sorrow</b>	04   06   02	06   07   06	01   01   01	04   01   03	01   00   00	09   01   03	<b>75   84   85</b>	00   00   00
<b>Surp.</b>	00   00   00	00   00   00	00   00   00	10   07   06	00   00   00	01   00   00	00   00   00	<b>90   93   93</b>

Figure 2.8: Confusion matrix with Random Forest using all eight emotions for subsets  $S3$  —  $S4$  —  $S5$ .

using  $S3$ . Furthermore, the Neutral expression calibration obviously increases the dissimilarity between other emotions.

Comparing the results obtained using the non-differential and differential subsets, in the latter case, it is possible to observe some improvements on the recognition of three particular emotions, Anger, Neutral and Sorrow. The increment of the recognition accuracy of the Neutral expression was expected due to the calibration that uses the

	Angry	Disgust	Fear	Joy	Sorrow	Surprise
Angry	<b>86   88   93</b>	03   05   01	00   00   00	00   00   00	10   07   05	00   00   00
Disgust	04   04   02	<b>94   92   96</b>	01   01   01	02   01   00	02   01   01	00   00   00
Fear	01   00   00	01   00   00	<b>86   88   91</b>	00   00   00	07   05   05	07   06   06
Joy	01   01   00	01   04   00	00   00   00	<b>95   96   98</b>	01   01   00	00   00   00
Sorrow	09   06   05	03   01   01	07   02   03	00   00   01	<b>82   91   90</b>	00   00   00
Surprise	00   00   00	00   00   00	08   08   06	00   00   00	00   00   00	<b>92   92   94</b>

Figure 2.9: Confusion matrix with Random Forest using 6 emotions (without neutral and contemptuous) for subsets  $S3$  —  $S4$  —  $S5$ .

Neutral facial emotion. The increment in the Anger and Sorrow emotions recognition accuracy was a consequence of the better recognition of the Neutral emotion since they were often mistaken as Neutral. However, we also noticed a decrease of accuracy for the Disgust expression recognition using the subset  $S4$ . In this case, the calibration reduced the Disgust dissimilarity in comparison with Fear, Joy, Sorrow and Surprise, resulting in misclassification towards Surprise emotions.

An interesting result about the classifier performances using subset  $S5$ , is that it has proved its capacity to exploit the best aspects from the two  $S5$  subset's components,  $S3$  and  $S4$  to improve the emotion recognition accuracy. For example, the classifier used  $S3$  features to avoid the misclassification of the Disgust expression, typical misclassification when using only  $S4$  features. More in detail, we report in Table 2.8 and Table 2.9 the confusion matrices obtained with Random Forests classifier using respectively eight and six (without Neutral and Contemptuous) emotions for subsets  $S3$  —  $S4$  —  $S5$ . For sake of brevity, we do not report the confusion matrices obtained for the two seven-emotion tests (eight emotions except Neutral), because they provide intermediate results between those achieved for eight and six emotions.

Analyzing the literature of the emotion facial recognition systems and comparing them with the obtained results reported in Table 2.6, we realized that the emotion recognition method based on our proposed features outperformed several alternative methods of feature extraction. We compare our method to:

- MPEG-4 FAPS [PB02], Gabor Wavelets [BLFM03] and geometrical features based on vector of features displacements [MEK03] methods with respect to the results obtained by Random Forests classifier using  $S3$ . These real time methods only classify the six universal facial expressions without using differential features with respect to *Neutral* face with an accuracy of 84%, 84% and 72%, respectively;
- three differential feature methods Michel *et al.* [MEK03], Cohen *et al.* [CSG<sup>+</sup>03]

and Wang *et al.* [WY07] with respect to the results obtained by Random Forests classifier using *S5*. Also these state of the art (SoA) methods allow the detection of only six universal emotions with average accuracies of 73.22%, 88% and 93%, respectively.

To summarize, in Table 2.7, we report the performance comparison between the aforementioned emotion facial recognition methods considering only the six universal emotions [EF75].

## 2.2.4 Discussion and Conclusions

We propose a flexible and novel geometric method that extracts facial features inherent to emotions from image stream captured by off-the-shelf hardware (e.g. webcams). The proposed method solves the four typical emotion recognition issues and allows a 94% of accuracy on emotion classification, which is higher than literature methods. Moreover, the method versatility allows the use of different facial landmark localization techniques (i.e. both marker-based and markerless) being a modular solution. However, it still requires that the face recognition technique presents as output a minimum number of landmarks associated to basic facial features, such as mouth, eyes and eyebrows.

Compared to traditional methods, our method allows, beyond the classification of the six universal emotions, the classification of two other emotions: Contemptuous and Neutral. Therefore, it can be considered as a complete tool that can be incorporated on facial recognition techniques for automatic and real time emotion classification of facial emotions. As proof of concept, we incorporated this tool in a LIFEisGAME [AMQO13, LIF09] game mode, where the user must match a target expression retrieved by the game. User's face is captured and emotion classified in real time. Regarding practical performance, we verified that it is more stable when we apply a neutral face calibration, classifying correctly the emotions expressed. However, it requires that the user knows how to make the expression properly. Problems regarding environment (background and lightning changes) were not addressed.

Thus, the adopted dimensionality reduction through extraction of eccentricity and linear representation of these 66 features combined with a Random Forests learning was enough to accurately recognize the universal emotions.

Results obtained in this study confirmed the impact of minimum facial features,

appearing as a subset of the 66 landmarks used during tracking. We used this knowledge during feature definition in Chapter 3 - Facial MoCap Tracking. In addition, feature extraction and machine learning knowledge was used as a baseline for MoCap VR methods in Chapter 5.



# Chapter 3

## Facial MoCap Tracking

*The essence of creating believable facial animation with MoCap is to capture the whole expression of an actor's face and reproduce it on a 3D character. This "expressiveness" extraction is only possible if the acquisition setup is able to capture every trait and characteristic of the performer [LYYB13b]. In this chapter, we deliver a detail state of the art overview of recent advances in the field of facial MoCap tracking. Analyzing literature limitations, we describe a novel method. For method's evaluation, we deployed two markerless facial MoCap systems (i.e. Saragih et al. [SLC11b] and Cao et al. [CHZ14]). In addition, we build two protocols that allow the evaluation and validation of generic CV algorithms. We also describe one of the protocols, FdMiee. FdMiee protocol was used to validate and compare the novel method with literature and, later, to visually test the VR methods of Chapter 5.*

---

**References:** Results from this chapter were presented at VERE European project consortium [VER10]. Facial MoCap Tracking poster can be found at Appendix C. The protocols for CV system validation resulted in a journal paper currently under review in the Journal of Pattern Recognition and Image Analysis (Appendix D).

### 3.1 Background

Convincing reproduction of human facial movements in 3D characters is one of the greatest challenges in CG. Due to the diversity and complexity of human face, high

quality animation through traditional methods (i.e. manual keyframe) hardly allow replication of the subtleties of real movements [Lew06]. To facilitate this process, MoCap systems have been widely used to capture real face movements and create facial animation in a process usually called performance-driven facial animation. This definition as performance-driven facial animation (i.e. MoCap facial animation) was first introduced by Lance Williams [Wil90] in 1990. As proof-of-concept, William’s [Wil90] proposed a static face scan and facial deformation was driven by a marker-based approach using video (2D). During that decade, MoCap was hardware-based and their usage to create short animations became ordinary. This led to an explosion in the development of simpler software-based approaches. Since that, literally hundreds of independent published works appeared. This chapter focus in the commonly used and recent approaches. For a more complete literature overview, we forward the reader to the survey [DN07].

Today, in facial MoCap tracking, we distinguish two kinds of approaches: **equipment-based** and **markerless**:

- **equipment-based:** These approaches have been widely used within entertainment industry due to their high fidelity results. They include marker-based [RE01, FL03, AM06, HCTW11], camera arrays [BHPS10, BHB<sup>+</sup>11] or structured-light projectors [ZH04, WLVP09] approaches. Besides being more intrusive, marker-based approaches are physically time consuming since they require an accurate placement of a large number of markers in actors’ face [Lew06]. These characteristics limit their usage to production environments (see example of Figure 3.1). Additionally, the markers are typically sparsely distributed in actor’s face requiring further manual corrections or physical deformation priors [BLB<sup>+</sup>08] to re-introduce fine scale dynamics. Consequently, equipment-based approaches are not suitable for general users because these equipments are only attainable by companies.
- **markerless:** These methods appear as a more practical and less intrusive solution for non-expert users as they usually rely on off-the-shelf acquisition hardware and do not require complex placement of markers [LYB13a]. As definition, markerless approaches locate semantic facial features (e.g. eyes, mouth, nose, etc) in video frames and, then, track them to trigger facial animation [CHZ14]. Initially, optical flow tracking was the mostly used, however, due to the high sensitivity to noise, it was considered not suitable to track fast movements [CWLZ13]. At that time, to overcome this robustness issue, several geometric constraints based techniques [EBDP96, PSS99, BV99b, CET01,



CXH03, VBPP05, CC08] started to appear, where we include the known Active Appearance Models (AAM) [CET01]. Nowadays, these techniques evolved to 2D facial shapes trackers, like CPR (Cascade Pose Regression) [DWP10, CWWS12, CWLZ13, CHZ14], CLM (Constrained Local Models) [SLC11b, SLC11a, AZCP13] or SDM (Supervised Descent Method) [XDIT13]. Recent comparison studies [CHZ14] showed that CLM have problems in subtle and asymmetric motion tracking, due to the limited local search radius, and require additional processing before being applied to 3D characters [SLC11b]. This problem also persists in SDM techniques [XDIT13]. In parallel, with the increase of RGB-D cameras availability at consumer-level (e.g. Intel RealSense, Microsoft Kinect, etc), depth-based methods were developed [WBLP11, BRM12, LYYB13a, CWS<sup>+</sup>13]. Compared to the majority of previous video-based approaches, RGB-D methods have more accurate and robust results, even for arbitrary users, i.e. without training and calibration [BRM12, LYYB13a]. In 2013, to reach RGB-D's accuracy and performance level, video-based techniques exploited 3D shape regressors (trained for each user) and applied them to 2D frames, retrieving 3D facial shapes [CWLZ13]. More recently, Cao [CHZ14] proposed a Displaced Dynamic Expression Regression model for face shape representation. This method is one of the most promising markerless approaches, since it removes the calibration step of [CWLZ13] and, simultaneously, outperforms previous video-based and RGB-D techniques (see Figure 3.2). However, all the previous systems rely in the usage of a face model training. The usage of predictive models leads to incompatibilities when leading with persistent partial occlusions (like in VR scenarios [CHZ14]) and no sensitivity to atypical facial movements not learned in the model's training.

Summarizing, today there are hundreds of facial MoCap trackers. Still, the adoption of facial models for capture lead to problems in the tracking of subtle and atypical movements resultant from the diversity of faces (i.e. morphologies and behaviors) not predicted by model's training. Another limitation arises in the presence of persistent partial occlusions created, as example, by VR hardware. We address the occlusion issue at Chapter 5 - MoCap VR Methods.



Figure 3.1: Marker-based MoCap example: Actor Mark Ruffalo in the role of Hulk at "*The Avengers: Age Of Ultron*" (2015). Source: <http://www.cosmicbooknews.com>

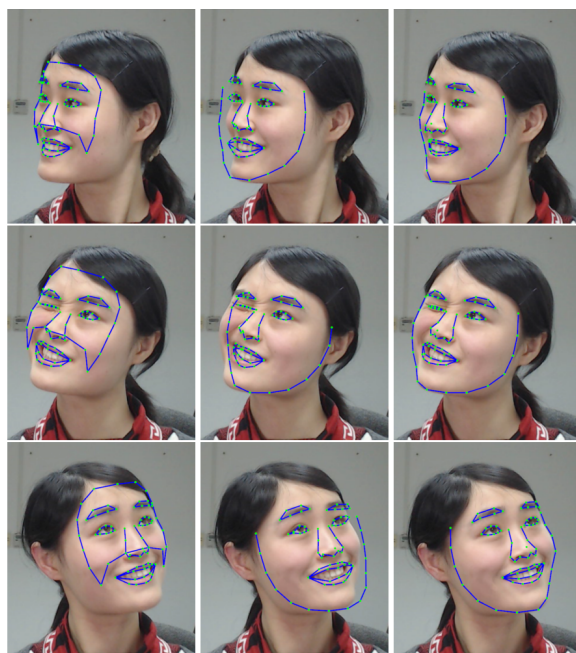


Figure 3.2: Markerless video-based approaches - tracking comparison (from left to right): (i) User-specific algorithm [CWLZ13], (ii) 3D CLM [SLC11b] and (iii) DDE user-free algorithm [CHZ14].

## 3.2 Methodology

This section describes the adopted methodology to create a novel method for facial MoCap tracking. Simultaneously, we proposed two novel databases protocols: FdMiee (Facial database with Multi input experiments and environments) and FACIA (Facial Multimodal database driven by emotional induced acting). FdMiee database is used for tracking method comparison and validation. Thus, we present an overview of FdMiee protocol, respective proof-of-concept database and protocol’s conclusions. For a complete description of both protocols, we suggest the reading of Appendix D.

### 3.2.1 Novel Method: Facial MoCap Tracking

Our novel method aims to capture the unique facial features of the user and overcome the following literature’s limitations: subtle and unique facial traits tracking, like asymmetric and cheeks’ movements. For simplicity, we define these unique features as uncommon movements/features. To reduce the expertise requirement, we adopt a markerless approach with capture using off-the-shelf hardware, i.e. webcams.

#### 3.2.1.1 Features Definition

In the MoCap Fundamental Science (Chapter 2), we state the minimum facial features that need to be tracked to define a static face shape properly. Although, observing literature facial models [SLC11a, LYYB13a, CHZ14] and our emotion recognition study (Chapter 2), to a proper movement description we need to track more features, specially in the mouth and eyebrows. But accessing and tracking subtle and uncommon movements of the face (e.g. cheeks movements) remains a challenge. As a result, in our method, we propose to combine facial features from Saragih’s *et al.* model of Figure 3.7 which compared to other literature models [LYYB13a] provided more features regarding face limits) with Optical Flow tracking [Far00, Far03] for additional subtle and uncommon facial features tracking (green dots in the Figure 3.9).

#### 3.2.1.2 Method’s Overview

Taking into account the features defined and this thesis goals, we adopt common webcams or Head Mounted Cameras (HMC) as capture hardware.

We start by answering the question: *why do we chose Optical Flow to track the face features?* In the Background section, it was stated that Optical Flow usage was abandoned due to its sensitivity to image noise, leading to error accumulation. Alternative model-based trackers suffer from the opposite problem: lack of sensitivity, not being able to track subtle and uncommon movements. Model-based methods result from learning processes that are not sensitive to uncommon movements (i.e. models only predict movements used in training procedures). Therefore, we conceive the following hypothesis:

*Optical Flow can be used to track subtle and uncommon movements*

In our method, we adopted the Farneback’s Optical Flow algorithm: ”Two-Frame Motion Estimation Based on Polynomial Expansion” [Far00, Far03]. To overcome the lack of stability of Optical Flow , we deployed two stabilization algorithms:

- **Baseline Movement Estimation (BME):** which uses image noise estimation and applies a real-time filtering technique.
- **Zone-based stabilization:** this algorithm ensures the stability of the face model landmarks during tracking. In our implementation, we used the face model structure from Saragih *et al.* [SLC11a] and store the model in a configuration file during *calibration*. We define the model structure using the relationship between facial zones (e.g. mouth, jaw, eyes, eyebrows, etc) and their relative position (i.e. orders vector), i.e. we create a hierarchy (see Figure 3.3). Therefore, to a certain face model, we define: zones (filled purple boxes at Figure 3.3) and landmarks (blue boxes at Figure 3.3). Each zone has an *id* and can contain the following childs: zones and landmarks. Note, landmarks cannot have zones as childs. If the child is a zone we store the following parameters: *id*, if it is *horizontal* or *vertical* and their *positioning order*. If the child is a landmark, we store: *id*, *type* and *order*. Figure 3.4 shows a simplified example of a model hierarchy with two zones (blue and purple boxes). The configuration file is only changed if we change the face model used. In the *Calibration* stage we introduce how the hierarchy is created. In *runtime*, we describe how we update the landmark positions using Optical Flow and how the hierarchy ensures that facial model structure is maintained (see *Runtime*).

As already mentioned, in the stabilization algorithms we apply the facial model from Saragih *et al.* [SLC11a]. Ahead, we present a fully description of methodology and

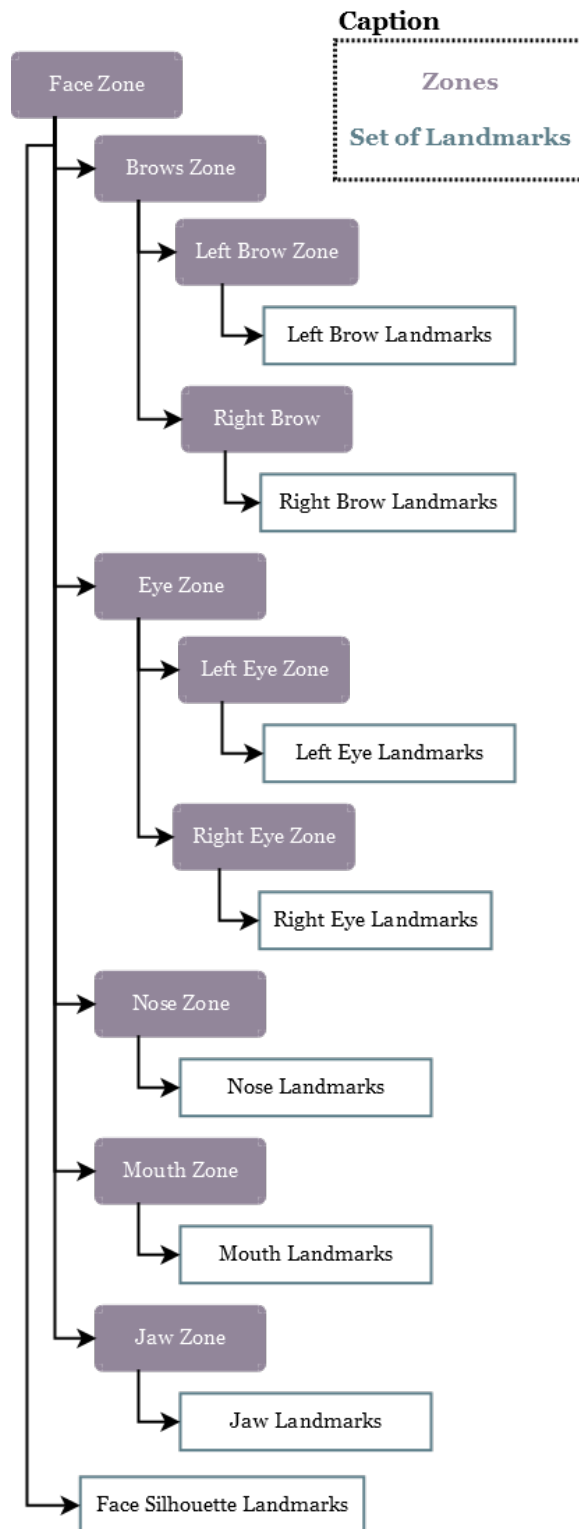


Figure 3.3: Hierarchy structure containing facial zones and landmarks used to setup the XML configuration file.

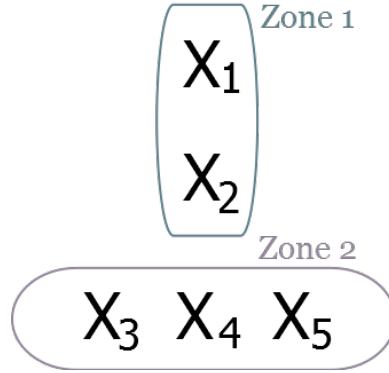


Figure 3.4: Zones and landmarks hierarchy example: The blue box is type Zone vertical, id 1 and position order. This Zone 1 has as childs two landmarks (type),  $X_1$  and  $X_2$  with id 1 and 2, respectively. The purple box is the Zone id 2, horizontal and contains the child landmarks  $X_3$ ,  $X_4$  and  $X_5$  with id 3, 4 and 5, respectively.

inherent algorithms. Figure 3.5 shows that our method is divided in three stages: *calibration*, *runtime* and *reset*.

### 3.2.1.3 Calibration

This step is only performed one time per user. As shown in the Figure 3.6, the *calibration* stage receives as an input the image stream from capture hardware and a configuration file. As an output, it returns a Region of Interest (ROI), containing the user's face, a BME and an updated hierarchy structure, required for the Zone-based stabilization.

The algorithms inherent to each output parameter are:

- **ROI** contains the region of the image where is located the user's face. The ROI limits the tracker search to increase efficiency. It results from the output of an optimized version of the haar cascade algorithm [LM02] restricted to single face detection.
- **BME** stores a measure of the residual image noise not related to facial movements. We ask the user to hold still for 5 seconds, run the Optical Flow and make an estimation of the average of movement at pixel level. We store the value in the BME parameter for further removal during *runtime*.
- **Hierarchy structure setup:** to create the hierarchy structure for landmarks stabilization, we run one time the Saragih's model-based tracker [SLC11a, SLC11b]. This step retrieves 2D basic features landmarks in user's face (see Figure 3.7).

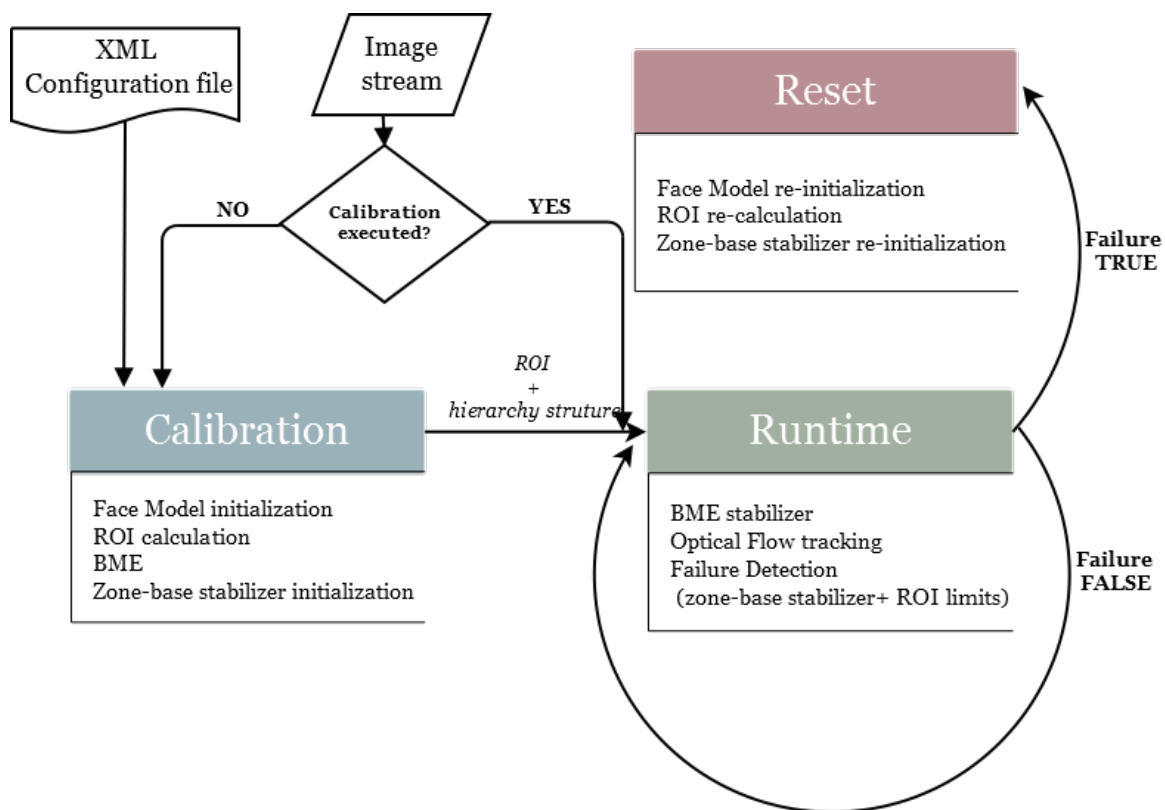


Figure 3.5: Novel method for facial MoCap Tracking: inputs in diamond boxes, method's stages (colored boxes) and respective sub-methods.

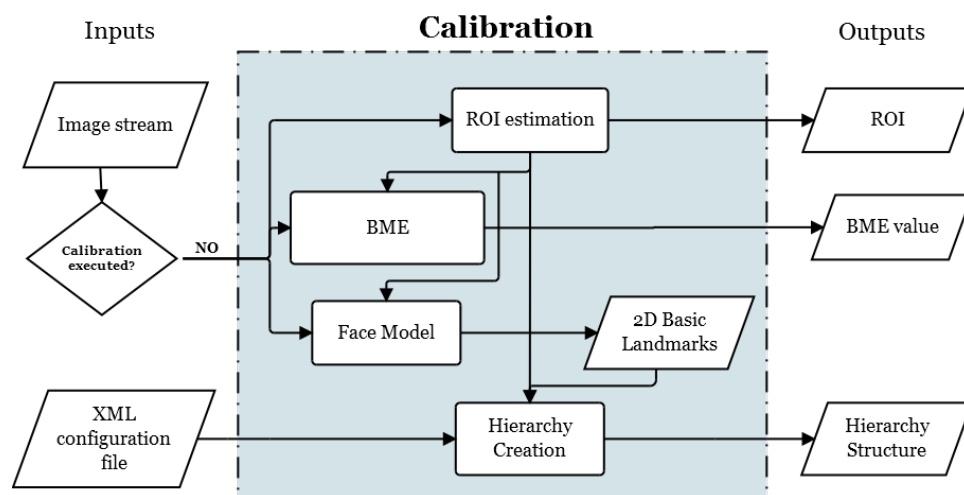


Figure 3.6: MoCap Tracking method - Calibration stage that receives the image stream and XML configuration file and returns the ROI, BME value and Hierarchy structure of the user.

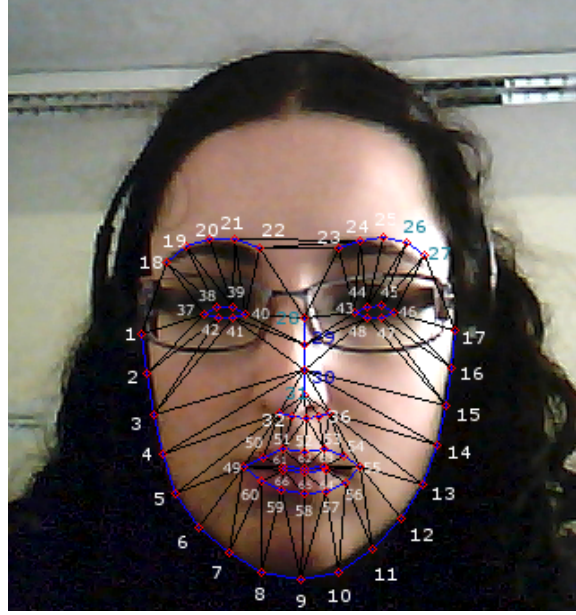


Figure 3.7: Face model landmarks of facetracker of Saragih *et al.* [SLC11a, SLC11b].

Combining those 2D landmarks positions with XML configuration file containing information of face zones of the model [SLC11a, SLC11b]) (Figure 3.3), we generate an user-specific hierarchy structure (Figure 3.8). The hierarchy structure stores: zones information; zones' current and last position in space; zones' limits (maximum between XMin, XMax and YMin, YMax in Figure 3.8 used to define the zones' ellipses); zones' influence radius (further used to define Optical Flow area of influence); siblings and childs (classified as zones or landmarks) with respective orders. Childs that are landmarks cannot have childs and, therefore, do not have zone limits information. As a result, the Figure 3.9 shows the landmarks and respective zones represented by ellipses (sharing the same color).

The calibration does not require any manual intervention from user. Regarding usage of HMC, we deployed algorithm for distortion removal of Go Pro Hero © videos that should be applied in this stage.

#### 3.2.1.4 Runtime

Figure 3.2.1.4 shows the group of processes that run in real-time. As an input, the *runtime* stage uses the image stream and outputs from *calibration*. As an output (per frame), it retrieves: a ROI and updated hierarchy structure (containing the updated 2D landmarks positions); the displacement map from the Optical Flow algorithm and a boolean from failure detection methods (see Failure detection section).



## Hierarchy Structure

<i>Parent</i>	
<i>Zone Type (TRUE/FALSE)</i>	
<b>Position Information</b>	<b>Last Position Information</b>
<i>X</i>	<i>X</i>
<i>Y</i>	<i>Y</i>
<i>XMin</i>	<i>XMin</i>
<i>XMax</i>	<i>XMax</i>
<i>YMin</i>	<i>YMin</i>
<i>YMax</i>	<i>YMax</i>
<i>InfluenceRadius</i>	
<i>Siblings</i>	
<i>Childs</i>	<b>Caption</b> <i>Zone Type</i> <i>Both Types</i>
<i>Orders Vector</i>	

Figure 3.8: MoCap Tracking method - Hierarchy structure template and parameters required. In the template XMin, XMax, YMin and YMax define the zone limits, where the maximum between the min and max of the coordinates is used to build the zone ellipse.

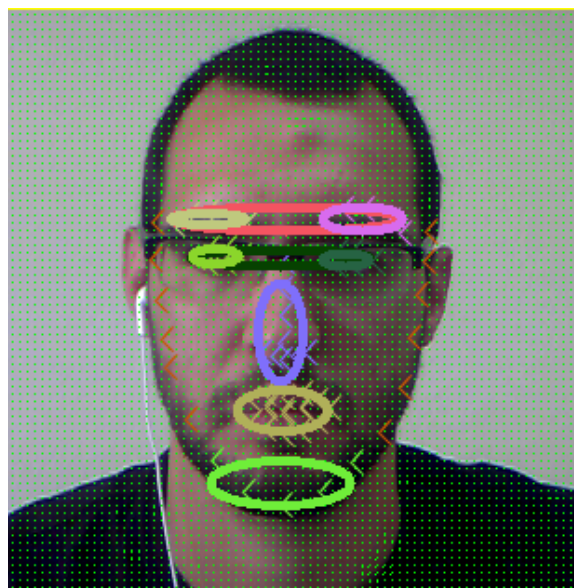


Figure 3.9: MoCap Tracking method - Visual result of hierarchy structure: landmarks belong to zones (i.e.ellipses) with the same color.

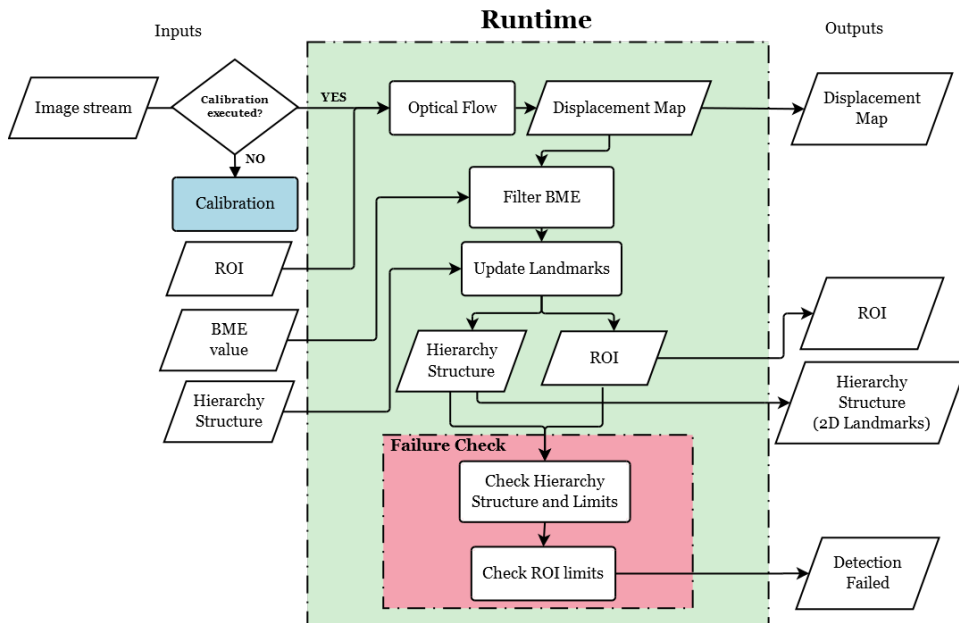


Figure 3.10: MoCap Tracking method - Runtime stage uses the outputs from Calibration and the image stream and run the Optical Flow combined with stabilization methods to update the 2D landmarks stored in hierarchy format and the ROI. This stage also contains a Failure Check.

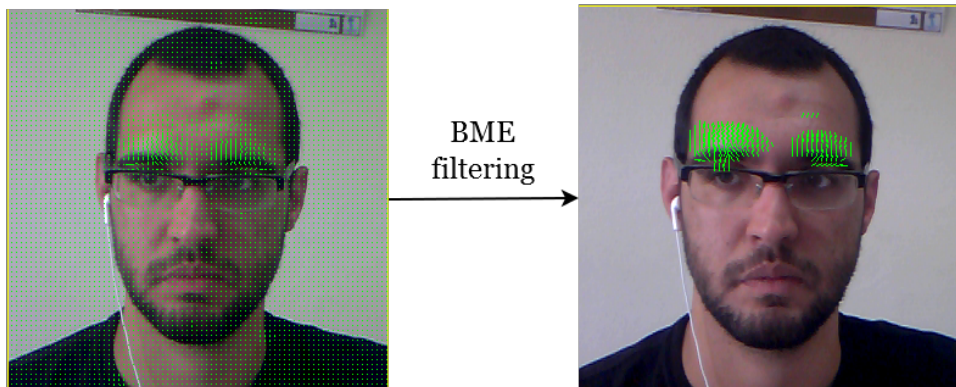


Figure 3.11: MoCap Tracking method - Optical Flow result, before (left image) and after (right image) BME filtering.

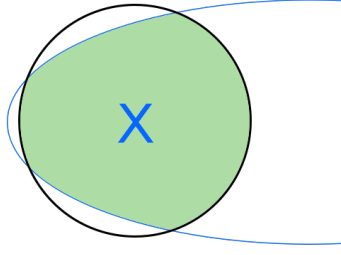


Figure 3.12: MoCap Tracking method - Landmark update using Optical Flow: the Optical Flow displacement that influence the position of the landmark X (green area) is defined by the interception between zone limits (blue ellipse) and circumference with radius defined by influence radius (black circle).

To update the 2D landmarks positions, we apply the BME, i.e. landmark displacements below BME value are ignored. Figure 3.11 shows the Optical Flow tracking before and after BME filtering. Using the filtered Optical Flow, we update the 2D landmarks positions as follows: we define each landmark as a center from a circumference with radius equal to respective radius of influence (assigned in the hierarchy structure). To update each center (i.e. landmark) position, we calculate the average displacement of pixels inside the area resultant from the interception between respective circumference and parent zone limits. As output, both 2D landmarks positions and hierarchy structure (i.e. positions and limits of each zone/landmarks) are updated. Figure 3.12 shows an example to one landmark.

To each update cycle, the ROI is also re-calculated using difference between the limits of the updated 2D landmarks positions and with previous respective limits. This *changing ratio* is used to scale and translate the ROI, accordingly.

Besides the face model landmarks contained in the Hierarchy structure, we retrieve, as methods' output, the displacement map calculated by the Optical Flow algorithm. The output map is delimited by the face model. The displacement map contains all the facial movements not described by the model (see green dots at Figure within the hierarchy structure), i.e. uncommon movements.

**Failure detection:** To ensure that the face is correctly tracked we evaluate two parameters: Hierarchy structure's limits and ROI's limits (red rectangle of the Figure 3.10). Hierarchy structure's limits show if the integrity of facial model is maintained. If an "incoherence" in a zone/landmark is detected (e.g. landmarks with sequential orders exchanged places), we revert all the landmarks to the last position stored (where

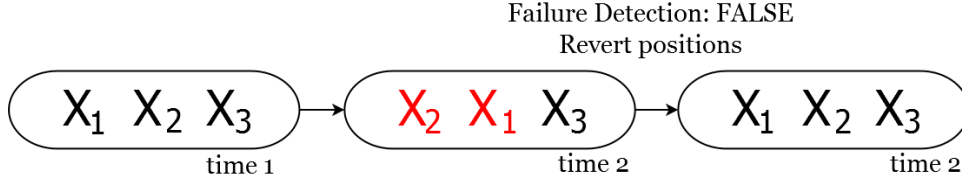


Figure 3.13: Failure detection when hierarchy structure of the facial model is not maintained: at time 1 we have the landmarks  $X_1$ ,  $X_2$  and  $X_3$  in the correct order. A wrong update at runtime lead to a change of  $X_1$  and  $X_2$  positions. Failure detection method detects the error and reverts the positions of the landmarks  $X_1$  and  $X_2$  at time2 to their correct positions at time1.

integrity was ensured by failure detection in the previous cycle). Figure shows an example of failure detection and how we revert landmarks to correct position with our method. The second parameter allows us to check if the ROI updated is inside the full image size. If not, it retrieves false to the failure detection variable. If it is true, it repeats the *update* stage of Figure 3.10. Everytime failure detection retrieves false as output, our method proceeds to *Reset* stage.

### 3.2.1.5 Reset

The *Reset* stage is summarized by Figure 3.14. The reset stage uses as input: the image stream, the ROI and hierarchy structure calculated before. This method forces the Saragih’s model tracker to re-detect the basic 2D landmarks and updates the hierarchy structure with the landmark new positions. As output, Reset retrieves a new ROI. The new ROI limits are defined by the basic 2D landmarks and the hierarchy structure with their positions and limits.

## 3.3 FdMiee’s protocol

We created two acquisition protocols to help the creation of facial databases. During protocols validation, we were able to acquire two proof-of-concept databases. The first database, FdMiee, captures faces under different environments conditions and expressions. FACIA extends FdMiee regarding expressions and emotions using RGB-D hardware, i.e. Microsoft Kinect and acquiring also audio during speech.

In this PhD thesis, we describe only the FdMiee protocol (definition, description, validation and results), since FdMiee is adopted to validate and compare our novel method with literature algorithms and, later, at Chapter 5, to test the method visually.

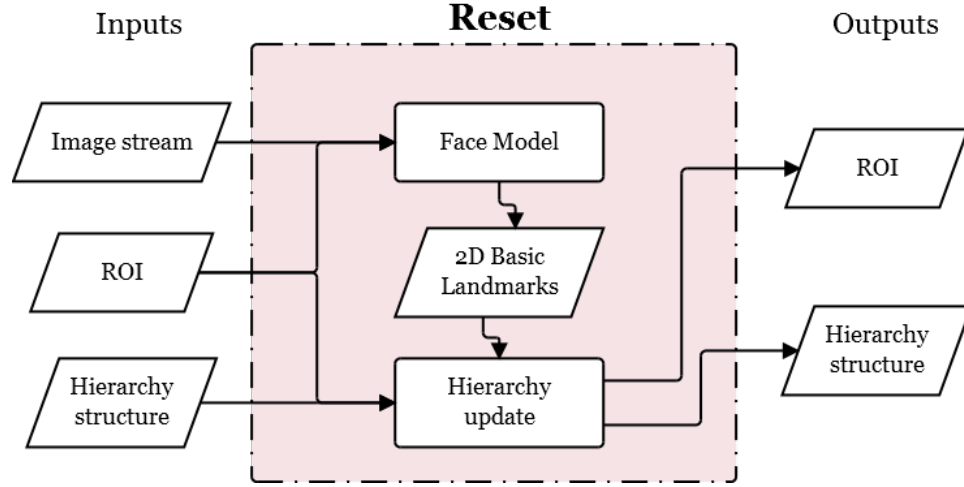


Figure 3.14: MoCap Tracking method - Failure detection resets the face model using the image stream and updates the ROI and Hierarchy structure.

The reader can read the full work, with both databases, at the Appendix D.

### 3.3.1 FDMiee: Protocol Methodology

Analyzing the background and details of facial data acquisition setups at Appendix D, we proposed to create a protocol through the characterization of three fundamental variables: *subject characteristics*, *acquisition hardware* and *performance parameters* (Table 3.1). These variables were classified as being either flexible or fixed, according to their role in the protocol guidelines. Subject characteristics and acquisition hardware are flexible variables, as they can be changed according to the system requirements. For example, use male subjects captured with a high-speed camera or other kind of hardware available, since they do not influence the guidelines of acquisition itself, but only interfere with the acquisition setup. In contrast, fixed variables such as performance parameters, influenced the definition of the guidelines, i.e. different performance parameters require us to take different steps for their simulation and acquisition.

*Subject characteristics* include gender, age, race, and other features that can be extrapolated from the subjects' samples. Subject variables introduce specific facial behaviors (e.g. cultural variations in emotion expressions) in the database. Regarding, *acquisition hardware*, we enabled the usage of any type of input hardware according

Protocol Variables		
Flexible		Fixed
Subjects Characteristics	Acquisition Hardware	Performance Parameters
Gender Age Race (...)	Webcam HD Camera Infra-Red Camera Microsoft Kinect High-Speed Camera (...)	External Parameters: Background Lightning Multi-Subject Occlusions Facial Parameters: Head Rotation Expressions: Macro Micro False Masked Subtle Speech

Table 3.1: Protocol’s flexible and fixed variables.

to acquisition specifications. Different combinations of the flexible variables can be applied to any of the fixed *performance parameters* guidelines. Performance variables describe the procedures for acquiring the data required for performance tests of CV algorithms. They are split into External and Facial categories, according to what we want to test. External parameters are related to changes in the environment, such as background, lightning, number of persons in a scene (i.e. multi-subject), and occlusions [Cot10, BKP05, BCL01]. These variables are almost infinite [HRBLM07a] due to their uncontrolled nature in real-life environments. Facial behaviours should contain facial expressions data triggered by emotions, such as macro, micro, subtle, false, and masked expressions [EF78, EF75, Mar09, BMT13] or even speech information. Ekman *et al.* [EF78] defines six universal emotions: anger, fear, sadness, disgust, surprise and happiness. These universal emotions are expressed in different ways according to a person’s mood and intentions. The way they are expressed leads us to an expressions-classification:

- **Macro:** These expressions last between half a second and 4 seconds. They often repeat and fit what is being said during speech. Facial expressions of high intensity are usually connected to six universal emotions [EF78, EF75, Mar09];
- **Micro:** Brief facial expressions (e.g. milliseconds) related to emotion suppres-

sion or repression [EF78, EF75, Mar09];

- **False:** Mirrors an emotion that is deliberately performed, and is not being felt [EF78, EF75, Mar09];
- **Masked:** False expression created to mask a felt macro-expression [EF78, EF75, Mar09];
- **Subtle:** Expressions of low intensity that occur when a person starts to feel an emotion or shows an emotional response to a certain situation, another person, or surrounding environment. This is usually of low intensity [BMT13].

Facial behaviors generated by speech usually contain a combination of the above expressions [KR12]. Following our methodology, we defined two protocols. To validate FdMiee protocol, we acquired data from eight subjects with different characteristics. We applied low-resolution, high-resolution, and Infra-red cameras as acquisition hardware variables. As performance parameter variables, we simulated multi-input expressions and environments to test the invariance and accuracy of facial tracking systems exposed to changes, e.g. different lighting conditions, universal-based and speech facial expressions. To validate the results, we executed 360 acquisitions and demonstrated the protocol's potentials to acquire data containing uncontrolled scenarios and facial behaviors. As a typical protocols' usage example, a research team has available database of 10 female subjects aged between 20-22. They would like to compile a database to test the head rotation tracking accuracy of the CV algorithm using a HD camera. Therefore, they define as *subject characteristics* the female gender and age range. Then, they choose a HD camera as *acquisition hardware* and afterward need to pick the Facial parameter: head rotation as Performance parameter. Finally, they followed our validated FdMiee protocol. Summarizing, in our protocols, we first choose the parameters to simulate as fixed Performance variables. So, we were able to define the acquisition guidelines. Secondly, we determine the hardware variable and generate an acquisition setup. Hardware variable is flexible, and thus changing it, does not impact in the guidelines. The same is verified using different subject characteristics.

### 3.3.2 FdMiee: Definition and Validation

Facial recognition and tracking systems are highly dependent on external conditions (i.e. environmental changes) [Bag12]. To reduce this dependency, we developed a

protocol based on our proposed methodology, for database creation with changes in terms of external parameters, such as light, background, occlusions, and multi-subject. For facial parameters, we setup guidelines to capture variation in head rotation, as well as universal-based, contempt and speech facial expressions. Table 3.1 summarizes the performance parameters acquired through the protocol.

### 3.3.2.1 Requirements

As protocol requirements, we setup the acquisition hardware and equipment to simulate the selected external and facial parameters.

### 3.3.2.2 Acquisition Hardware

The chosen acquisition hardware simulates realistic scenarios captured using three types of hardware. To test the protocol guidelines, we chose the following equipment: Low-Resolution (LR) camera; High-Resolution (HR) camera; Infra-Red (IR) camera.

The first two cameras (LR and HR) allow us to study the influence of image resolution on tracking, face recognition, and expression recognition [Tia04]. The IR camera allows to disregard lighting variation [WSE03, JA09, SGBP04] and provides a different kind of information than HR and LR cameras. The hardware used in FdMiee protocol should be aligned with one another to ensure future comparison between data acquired with different hardware.

### 3.3.2.3 Environment-Change Generation Equipment

To generate data with the defined parameters, we set the following environment elements:

**Background** A solid color and static background ease the process of detecting facial features and extracting information from the surrounding environment. The background should ideally be black (or very dark) to prevent interference with the IR camera (black color has lower reflectance compared to lighter colors)

**Lighting** The room must be lit up by homogeneous light, and not produce shadows or glitters in the subject's face. By taking these measure, we ensure that the skin color will have no variation throughout the acquisition process.



### 3.3.2.4 Protocol Guidelines

The subject sits in front of the acquisition hardware. The hardware setup is composed by three cameras (LR, HR and IR). The subject's backdrop should be black with some space between them, to have the possibility to move objects or subjects behind the main scene.

To perform the acquisition, we suggested the presence of two members: one to perform the acquisitions (A) and the other to perform environment variation (B). The subject sits in front of the computer monitor and one of the team members aligns them with the cameras. During the entire acquisition procedure, the subject should remain as still as possible, to avoid producing changes during the various acquisition procedures.

Before starting the experiment, each subject has access to a printed copy of the protocol. This reduces the acquisition time, since the subject already knows what is going to take place during the experiment. Each performance parameter simulated and introduced in the scenario has its own guidelines:

**Control** Team member A takes a photo with the subject in the neutral face.

**Lighting** Team member A takes 3 photos with different exposures (High, Medium, Low). This variable was only acquired in HR camera, because it is the only where it is possible to change the exposure level.

**Background** Team prepare the background to the acquisition.

1. Team member A starts recording;
2. Subject stay still during 5 seconds while team member B performs movement if necessary (only case of dynamic background);
3. Team member A stops recording.

**Multi-Subject** While subject is being recorded, team member B appear in the scene during 10 seconds.

**Occlusions** For total occlusion, subject starts in the center of the scene and slowly moves to a point out of the scene. For partial occlusions simulation, we take a photograph with a plain color surface, like a piece of paper covering the following parts of the face: Top, Left, Bottom and Right.

**Head Rotation** For each head pose (Yaw, Pitch and Roll), subject performs the movement in both directions while he is being recorded through the complete movement.

**Universal-Based Facial Expressions, plus Contempt** Subject repeats during 10 seconds the following emotion expressions, starting from the *neutral* pose to a full pose: *joy, anger, surprise, fear, disgust, sorrow* and *contempt*.

**Speech Facial Expressions** The subject reads a cartoon or text and is encouraged to express his feelings about it.

**Obtained Outputs** This protocol generates the following output data: HR and LR Photographies (.jpeg); LR camera videos - 15fps (.wmv); HR camera videos - 25fps (.mov); IR camera videos - 100fps (.avi).

The emotions generated through variation of facial parameters contain a combination of macro and micro (i.e. subjects can be repressing and suppressing feelings) as well as false (i.e. subject is making an effort to express certain emotions) and subtle (i.e. when subject cannot generate a high intensity expression) plus speech-based expressions.

### 3.3.3 Protocol Validation

Following the protocol guidelines, we acquired data from eight volunteers with the following subject characteristics: **Gender** Male/Female; **Glasses** With/Without; **Beard** With different formats/Without; **Age** 20-35 years

Figure 3.15 shows sample results from some of the performance parameters with the different acquisition hardware.

### 3.3.4 Protocol Conclusions

We presented a methodology to facilitate the development of two facial data acquisition protocols. Following this methodology, we create the protocols for simulation and capture of real-life scenarios and facial behaviors. To validate the protocols, two proof-of-concept database were captured: FdMiee and FACIA. To get detailed information regarding FACIA protocol, results and validation, the reader is forward to the full article at Appendix D. Generated databases can be used in a variety of applications, such as CV systems evaluation, testing, and training [Bag12]. Adopting

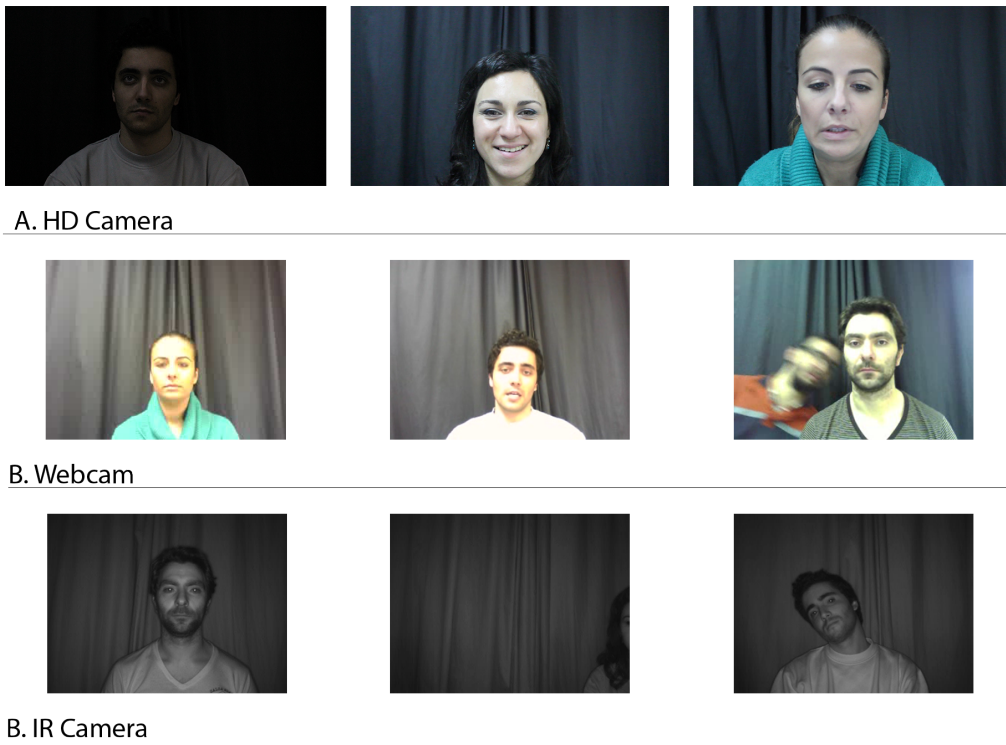


Figure 3.15: FDMiee samples results for HD Camera (A), Webcam (B) and IR Camera (C)

our methodology and protocols, we reduced the time required for customized database acquisition.

Throughout the protocol definition, we characterized two groups of variables: flexible variables (subjects' characteristics and capture hardware) and fixed performance variables (external and facial parameters). The FdMiee protocol focused on external parameters' simulation as a variation of the fixed performance. Thus, the protocol allows the acquisition of a facial database containing a large number of fixed parameters' variation (external and facial): lightning, background, multi-subject, occlusions, head rotation, universal-based, and speech facial expressions (Table 3.1). To validate the protocol, we performed an acquisition on eight subjects with different subject characteristics, leading to the creation of FdMiee database. FdMiee contains facial behaviors under different environment contexts. Hence, this protocol enables the generation of databases that are useful for a wide range of CV systems performance tests. Sample results were applied for facial MoCap tracking method's validation and to test visually the MoCap VR methods at Chapter 5.

### 3.4 Results and Discussion

The method implementation was performed using C++ language with OpenCV 2.4.9 GPU algorithms [ope14]. For comparison purposes, we deployed two literature algorithms: Saragih *et al.* [SLC11a] and Cao *et al.* [CHZ14]. This section shows and compares the results obtained with our method and deployed literature algorithms. As evaluation data, we use the FdMiee sample database due to the wide range of environment and expressions simulated and captured.

We checked if the tracker fails with environment changes (i.e lightning, background, multiperson) and tested methods' sensitivity to a wide range of face behaviors (i.e. poses and expressions). Our analysis tested if the algorithms fail in face features detection under these environment and behaviors variation. The output from the test was binary: face features were tracked or not. If algorithm failed the face detection under environment changes (0), if the algorithm tracked properly the face (1). Regarding facial behaviors: (0) if algorithm was not sensitive to face features behaviors and (1) for the opposite situation. Previous procedure was repeated to the six participants and we calculated a percentage of failure to each parameter (e.g. how much the algorithm fails in low lightning environments, the insensitivity to track subtle expressions during speech, etc).

In the Table 3.2, we present the percentage of failure for the three algorithms under different environment scenarios.

Environment Parameters ( % Percentage of failure)			
Parameters	Saragih <i>et al.</i> [SLC11a]	Our approach	Cao <i>et al.</i> [CHZ14]
<b>Light</b>			
Low	67	<b>33</b>	<b>33</b>
Normal	0	<b>0</b>	<b>0</b>
High	33	33	<b>17</b>
<b>Background</b>			
White	17	<b>0</b>	<b>0</b>
Static features	<b>0</b>	33	<b>0</b>
Dynamic features	83	<b>17</b>	<b>17</b>
<b>Multiperson</b>			
Static	83	83	<b>0</b>
Dynamic	67	<b>33</b>	<b>33</b>

Table 3.2: MoCap Tracking method - Percentage of failure in face detection under different environment changes. Results comparison between (i) Saragih *et al.* [SLC11a], (ii) Our facial MoCap Tracker and (iii) Cao *et al.* [CHZ14].

Analyzing the Table 3.2, we observe that our approach presents lower failure rate than Saragih *et al.* [SLC11a], with exception for static features. The higher failure rate at static features test is related to a failure in ROI calculation. The features behind the user produced confusion misleading the face detection making impossible a proper tracking. The same occur to static Multiperson scenarios, where the Saragih *et al.* algorithm also failed to track. This limitation can be eliminated with adoption of HMC to capture participants' face. In the Figures 3.16, 3.17 and 3.18, we show examples of tracking under different environment scenarios using our approach. These results show that proposed stabilization algorithms are suitable for Optical Flow, making the method stable enough to detect faces under real-life scenarios. In addition, our method outperformed the state of the art algorithm [SLC11a]. However, Cao *et al.* [CHZ14] presents the lower failure rates. Cao *et al.* method's performance was predicted, since his facial model was produced using a complex learning process with one of the most complete facial databases in literature [HRBLM07b].

Regarding behavior parameters variation, the percentage of failure reflects algorithms' accuracy detecting the face features when the participant changed his facial expressions. We defined that the algorithm "fails" when the tracker was not accurate enough to detect these features. The Table 3.3 shows the failure rate to the three algorithms

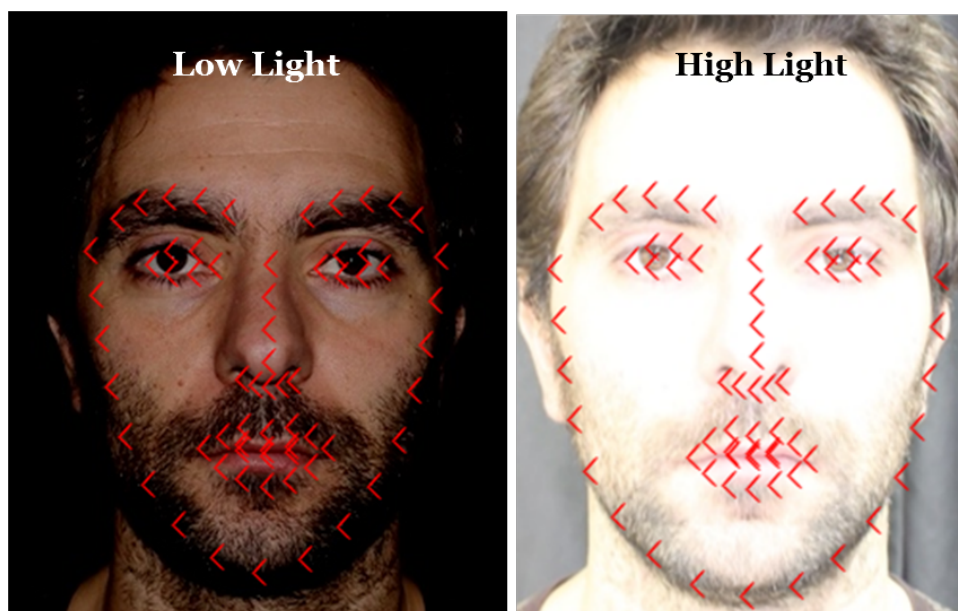


Figure 3.16: MoCap Tracking method - Environment light variation - Low (left) and High (right) conditions.

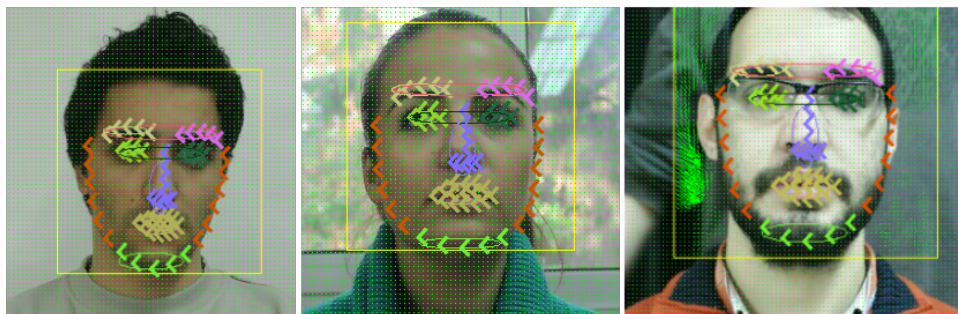


Figure 3.17: MoCap Tracking method - Environment Background variation - White (left), Static features (middle) and Dynamic features (right).

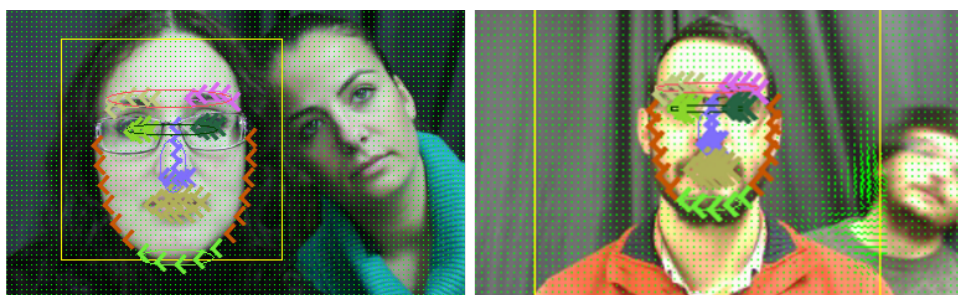


Figure 3.18: MoCap method - Environment with Multiperson - Static (left) and Dynamic person (right).



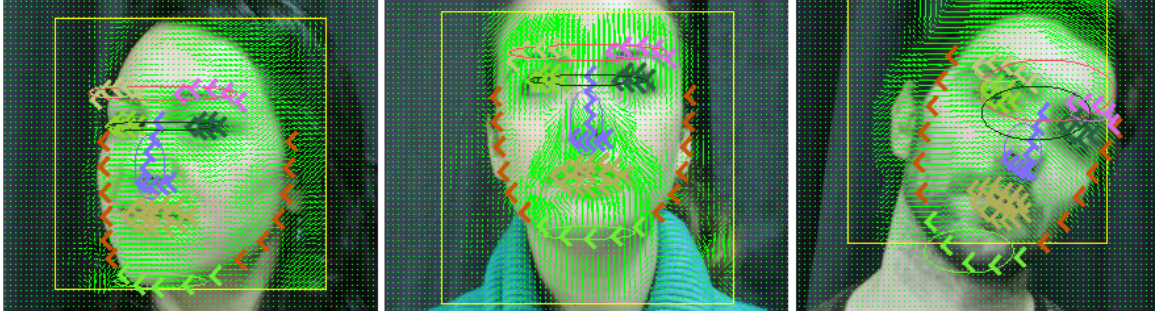


Figure 3.19: MoCap Tracking method - Behavior parameters: Head moving: Yaw (left), Pitch (middle) and Roll (right).

deployed.

Behaviors Parameters ( % Percentage of failure)			
Parameters	Saragih <i>et al.</i> [SLC11a]	Our approach	Cao <i>et al.</i> [CHZ14]
<b>Poses</b>			
Yaw	<b>17</b>	67	33
Pitch	33	<b>17</b>	33
Roll	<b>50</b>	67	<b>50</b>
<b>Expressions</b>			
Macro	25	25	<b>19</b>
Subtle	67	17	<b>0</b>

Table 3.3: MoCap Tracking method - Percentage of failure in detection subtle and macro expressions of different participants (i) Saragih *et al.* [SLC11a], (ii) Our facial MoCap Tracker and (iii) Cao *et al* [CHZ14].

The results show that our algorithm performs worse than other algorithms when the participants' face are not frontal to the camera. In general, tracking failure results from errors cumulation. In contradiction, to face expressions the novel method presents higher accuracy than Saragih's *et al.* tracker [SLC11a] (see left image of Figure 3.21). Even compared to Cao *et al.* algorithm (Figure 3.21 - right), analyzing the information retrieved from Optical Flow (green dots), our method is able to track uncommon features in the mouth, cheeks' wrinkles and eyebrows (see Figure 3.20 left). This information is not retrieved from any of the studied literature algorithms.

Similar results are observed regarding asymmetric features tracking. In the Figure 3.19 - Left, we show two examples of novel method, tracking: asymmetric features at the mouth (left) a joy expression (right).

More specifically, we point out two types of failure in the literature MoCap methods



Figure 3.20: MoCap Tracking method - Behavior parameters: Expressions: Disgust (left) and Joy (right). Red arrows represent the face model tracking and green points to the Optical Flow.

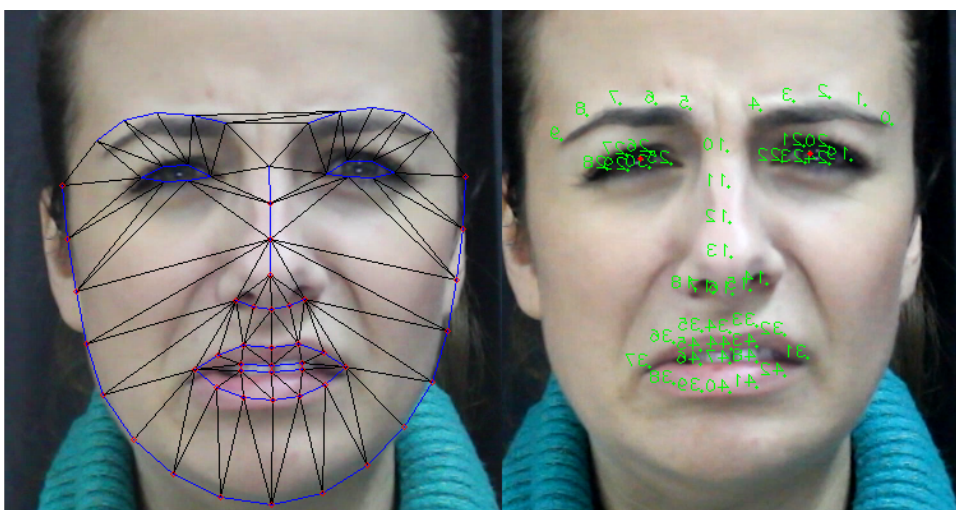


Figure 3.21: Comparison - Behavior parameters: Expressions: Disgust detected by Saragih *et al.* [SLC11a] (left) and Cao *et al.* [CHZ14] (right).





Figure 3.22: Cao *et al.* method [CHZ14] - Environment light variation - Low light with failure (left) and High light with no failure (right) detection.



Figure 3.23: Cao *et al.* method [CHZ14] - Environment with dynamic Multiperson scenario - test case with correct tracking (left), but after moving, the tracker fails detection (middle) or the detection jumps to the secondary participant (right).

deployed: (i) no detection of a face at all or (ii) wrong detection of facial features. The first type was the most frequent. Since (i) does not provide any information (i.e. appears as the original frame), we provide only two examples of this situation: right image of Figure 3.22 and middle of Figure 3.23. The second type is not so frequent, but occurred differently. Figure 3.22 shows one of the situations. The tracker started to detect the face properly (left), but after it jumped to the other person (right). Also, the wrong tracking can be originated from participants' head movements. Figure 3.25 show the head movements' failure to Saragih's *et al.* [SLC11a] method, and Figure 3.24 to Cao's *et al.* [CHZ14] method.

Summarizing, the novel method allows the accurate face features tracking, overcoming literature approaches in the detection of uncommon features, like cheeks movements. In general, to both environment and behavior variation, Cao *et al.*'s [CHZ14] algorithm outperforms our method. As mentioned, Cao *et al.* [CHZ14] work relies in a complex training using the Faces In the Wild database [HRBLM07b]. The Faces In the Wild

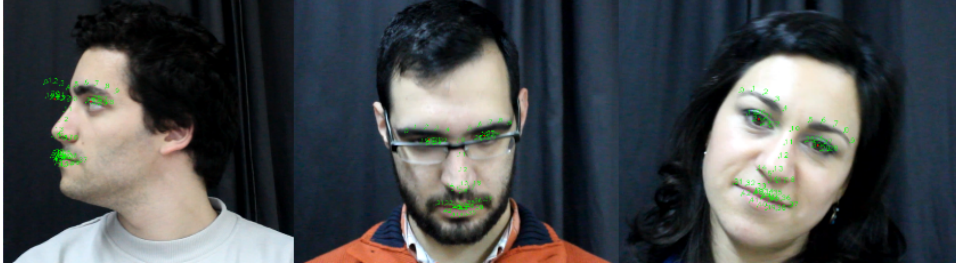


Figure 3.24: Cao *et al.* method [CHZ14] - Examples of failure in tracker's detection during the change of facial behaviors - Head moving: Yaw (left), Pitch (middle) and Roll (right).

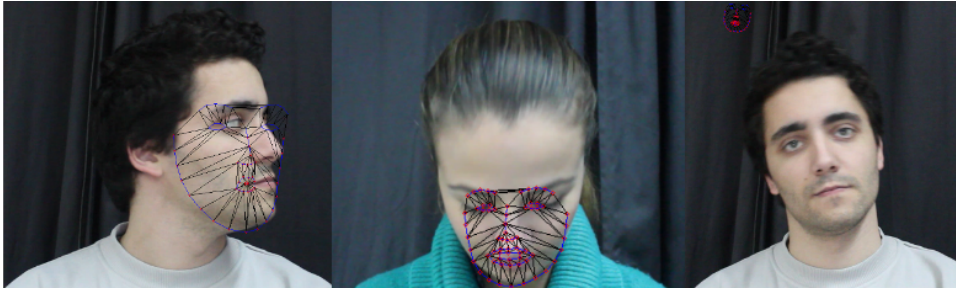


Figure 3.25: Saragih *et al.* method [SLC11a] - Examples of detection failure during the change of facial behaviors.

database is one of the most complete databases in literature, enabling the creation of high performance face models [CHZ14] that fit extreme environment scenarios, like low lightning (Figure 3.22 - right) and detection of asymmetric behaviors (Figure 3.21-right).

### 3.5 Conclusions

We propose a novel method for facial MoCap tracking that allows the tracking of user's facial movements using common RGB cameras (e.g. webcams). Moreover, our facial MoCap accesses characteristic and uncommon facial traits, such as cheeks' movements, derived from diversity of faces [McC93], without any prior knowledge of them. Our MoCap method applies an Optical Flow implementation combined with a noise filtering and a zone-based stabilization. Furthermore, we create two acquisition protocols for generic CV algorithms development, testing and validation. From these protocols resulted two sample databases and a journal publication (Appendix D). One of these two databases, FdMiee is used to validate and compare novel tracking method with literature approaches: Saragih *et al.* [SLC11a] and Cao *et al.* [CHZ14].

Analyzing the results we conclude that, in general, our method outperforms Saragih *et al.* [SLC11a] under both environmental and behaviors situations. Compared to Cao *et al.* [CHZ14], novel method presents higher failure rate under the tested environment situations. However, regarding the diversity of the facial features tracked, the novel method retrieves additional information about subtle features and face deformation beyond facial model landmarks, like cheeks and forehead movements. In the future, it could be possible to setup and deploy improvements in the stabilization methods to increase tracking accuracy under extreme environment scenarios, such as moving backgrounds or high intensity lightning.



# Chapter 4

## MoCap Facial Animation

*To reproduce movements of the face in 3D characters is a most challenging and time consuming tasks in CG. In the chapter 3, we explore one of the technological solutions to "partially" automatize the facial MoCap process. "Partially" because the MoCap systems only provide the input information to generate the animation. This input still needs to be "projected" to the 3D character's space and be correctly mapped to character's control structure to trigger deformation and produce animation. Due to space and topology differences between the user's and the 3D character's face, the mapping process is not trivial [BP14]. Hence, in this chapter, we introduce the facial animation pipeline concept. Then, we overview MoCap facial animation literature and pinpointed the problems that remain unsolved. As a solution, we deploy a novel methodology and evaluate the method's performance reproducing user's facial expressions captured by two facial MoCap tracking algorithms.*

---

**Reference:** Similarly to previous chapter, the results obtained using our mapping method were presented in: VERE and GOLEM EU projects' consortium meetings [VER10, GOL], LIFEisGAME National project as a game mode [AMQO13, LIF09] (see poster in the Appendix E). The mapping method is used by Porto Interactive Center (PIC)'s AVATAR pipeline to create real-time animation and was presented at Business Ignition Program (BIP) and iUP25k competition.

## 4.1 Background

Besides the fast developments lived today in CG, it is curious how many of these recent techniques arise from principles introduced during the decade of 70. The first 3D animation was produced by Frederik Parke in 1972 [Par72] and, since then, many approaches have emerged. To understand how these technologies are developed to improve traditional facial animation pipelines, it is relevant to have a clear idea of the process itself. Thus, we provide the basic concepts of facial animation pipeline and underlying key works. Then, we make the bridge to MoCap facial animation, where we explore the recent advances and problems that remain unsolved.

### 4.1.1 Traditional Facial Animation

The basic concepts of animation were introduced into CG pipelines by Lasseter in the 80's [Las87]. Still in the same decade, Degraf [Deg88] created the *Mike the Talking Head*, which was the first animated character that could interact with a person in real time. With these two approaches, two distinct roles appeared in CG pipelines: the artist and the technician, being the distinction between them not so clear. More specifically, looking through nowadays' animation pipelines of videogames and film industries, we distinguish four main stages: concept design, modeling, rigging and animation [OBP<sup>+</sup>12].

The pipeline starts with the concept design. In this step, artists start by defining the characters in scene, i.e. how they look like and how they should move. This visual information is further used by 3D artists to start modeling and creating characters' geometry [SSMCP02]. However, to manipulate character's geometry it is needed a intermediate control structure called rig. Rigging is defined by literature as: "*the process of taking a static, inanimate computer model and transforming it into a character that an animator can edit frame-by-frame to create motion*" [FMV<sup>+</sup>04]; and *the system engineering process that allows surface deformation* [MS07]. To create the rig, the rigger needs to have a high level of expertise and interact with both modelers and animators to understand what facial behaviors are expected. Due to uniqueness and the diversity of faces, the difficulty and complexity of this process climbs really fast, being one of the biggest challenges in the animation pipeline [OBP<sup>+</sup>12]. After rigging, can animators manipulate each control in the rig's interface and create the animation. Even with technological advances, many pipelines still produce animations manually (i.e. keyframe animation). In those situations, the quality of results rely only on

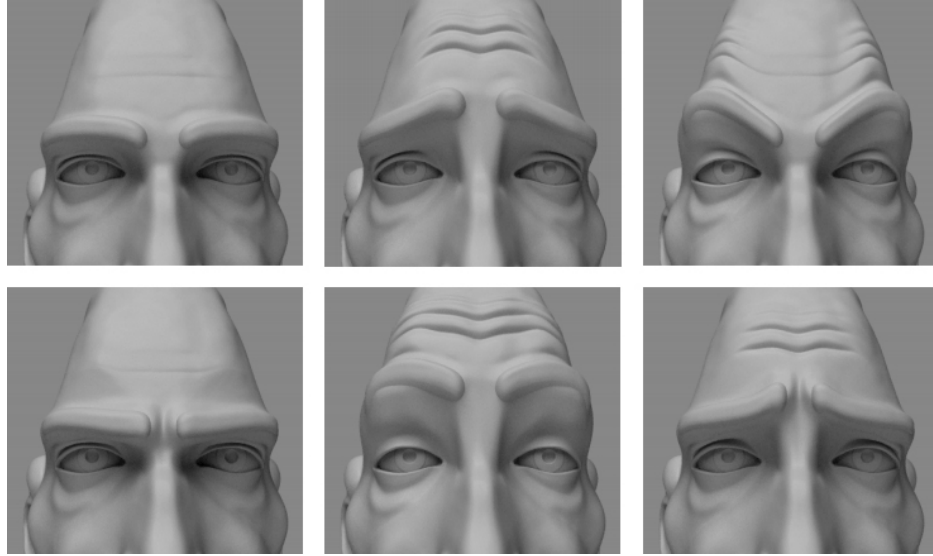


Figure 4.1: Blendshapes’ rig example: different poses of upper part of the face [SSMCP02].

animator’s skills and expertise, that needs to adapt to the type of rig and visual style of character [Sco93].

In general, we can find the following rig approaches: blendshapes interpolation, bone-based or hybrid methods. Hybrid methods are the combination of both blendshapes and bone-based. To create a blendshapes rig (see Figure 4.1), the rigger uses a group of meshes with the same topology and through sculpting generates various facial poses, i.e. shapes [Mar03]. The weighted interpolation between several shapes creates animation. As example, interpolating a face shape with eyes open with another with eyes close, creates a blinking animation. It is easy to understand why to build a blendshape-based rig is time consuming: the rigger needs to create each one of the face shapes and define their weight in the interpolation to produce each character’s facial behaviors. So, in the presence of a blendshape rig with few shapes, the range of animations that can be generated is limited. Moreover, the rigger should repeat the ”shape setup” to each character in the scene, increasing significantly the production time. Several methods were proposed to reduce the blendshape rig creation time, allowing the transfer of shapes between different characters in a process named retargeting [Orv07, DMB08, DMOB10]. For a complete state of the art in blendshapes topic, we forward the reader to the survey [LAR<sup>+</sup>14].

Bone-based rigs (see Figure 4.2) require the placement of controls (i.e. called articulation points/bones/joints) and the bind of the controls to the 3D geometry (i.e. skinning [LCA05, YZ06]), producing a highly articulated facial skeleton [OBP<sup>+</sup>12].

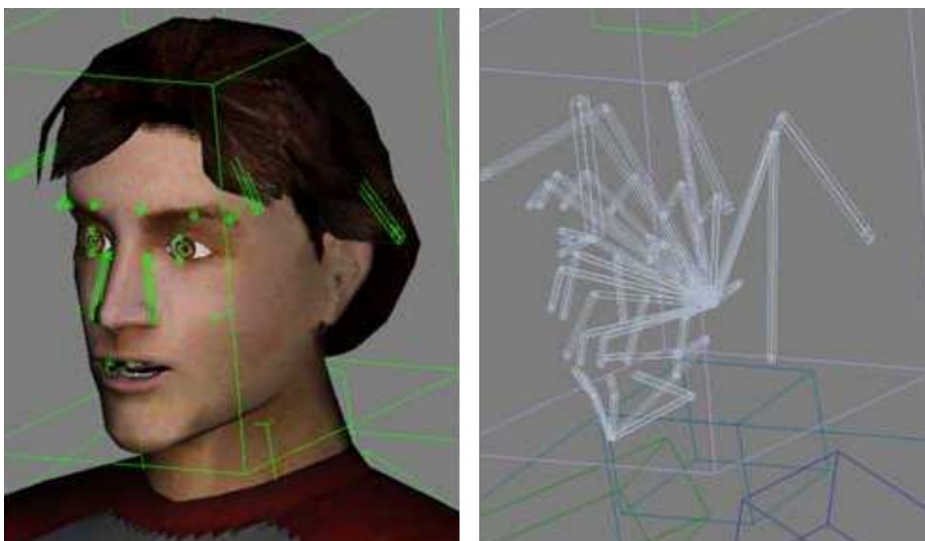


Figure 4.2: Bone-based rig example with Unreal game engine. (*Copyright 2001-2007 Epic Games*)

When creating a bone-based rig, the rigger: (i) places the controls; (ii) defines how they bind to each mesh's vertex and (iii) setup the influence area of each control using weights to produce believable deformations [Kom88]. Compared to blendshapes, the bone-based rig allows smoother results not limited by pre-defined poses. However, building a bone-based rig tends to require longer and more intense preparation, since each vertex is only animated by the bones around it [War04]. Like many others industrial processes, in facial animation we need to optimize two parameters: time and quality. For smoother and flexible results decreasing preparation time, researchers have adopted hybrid methods for rig creation, combining shapes with skeletal approaches [LCF00, LA10, LAR<sup>+</sup>14].

Thus, animation results depend upon rig's characteristics and structure. Rig's role can be clearly understood observing the data-flow of Figure 4.3, and for a better understanding of rigging process we suggest the reading of Orvalho *et al.*'s survey [OBP<sup>+</sup>12].

To get a more complete analysis of the traditional facial animation pipelines and their evolution throughout the years, the reader can access the surveys [NN98, DN07].

### 4.1.2 Facial Animation with MoCap

Markerless MoCap systems have been widely used to provide non-intrusive input motion data from real movements, avoiding artists to create animations from the



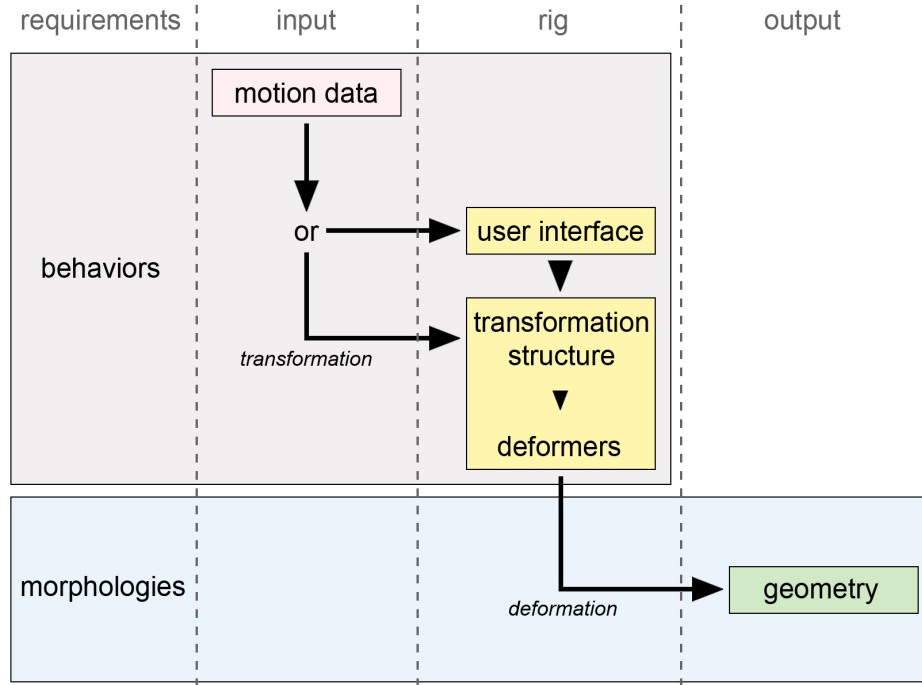


Figure 4.3: Rig data-flow structure. Stages: (i) requirements; (ii) input motion to activate the rig; (iii) geometry deformation. [OBP<sup>+</sup>12]

scratch [Lew06]. In the Figure 4.3, we observe the position of MoCap input data within the rig data-flow structure. Though, the usage of these CV algorithms (Chapter 3 - Facial MoCap Tracking) is not trivial, since the user’s face and 3D character differ in both shape and behaviors. To activate properly the rig, user’s movements need to be processed, transformed and filtered by a mapping method.

The commonly adopted solutions use example-based algorithms applied to the capture. Example-based algorithms require a calibration step with user’s expressions and, then, through 3D geometry and 2D texture registration, they approximate movements tracked to 3D character’s rig blendshapes (i.e. using statistical analysis to match movements tracked and character’s pre-defined shapes) [WBLP11]. Therefore, in example-base algorithms applied to capture, the MoCap system is not modular, being completely attached and tuned to a mapping algorithm. The no-modularity leads to more stable 3D character’s response to user’s facial movements. Yet, the stability is reached compromising the variety of movements that the 3D character is able to reproduce, which is limited by the number of blendshapes created. To automatize and improve the blendshapes rig, Digital Ira methodology [ARL<sup>+</sup>09, AFB<sup>+</sup>13, vdPJD<sup>+</sup>14] uses example-based rigs [LWP10] (see Figure 4.4). Here, the shapes in the rig are calculated from user’s expressions. Consequently, movements in animation are closer to user’s movements. The unquestionable quality improvements are introduced using,



Figure 4.4: Lightstage - acquisition hardware used in Digital Ira [DHT<sup>+</sup>00, AFB<sup>+</sup>13, vdPJD<sup>+</sup>14].

not only an expensive acquisition hardware (i.e. Lightstage) [DHT<sup>+</sup>00], but also with long and tedious calibrations with a long list of user’s expressions and scans.

Besides the unpractical usage of this kind of hardware (i.e. high costs), the calibration step is not suitable for non-expert users [CHZ14]. As a result, further advances focused in decreasing the number of expressions scanned [LYYB13a] for example-based capture and in creation of calibration free approaches [CHZ14], using complex statistical models generate through long learning preprocessing.

## 4.2 Methodology

Following the goal of this PhD research combined with literature challenges, we propose:

*to create a mapping method that adapts to user-choice MoCap tracking algorithm and reduces user-dependent calibration requirements.*

Compared to the approaches described in the Background section, the definition of the our mapping function is even more challenging, because: we do not apply example-based neither to capture nor rig’s; the function has to be free from any statistical information or learning processes from user, MoCap or rig; we intend to deliver a function that is independent from the facial MoCap tracking method, i.e. independent



Figure 4.5: Digital Ira results [AFB<sup>+</sup>13, vdPJD<sup>+</sup>14].

of number of landmarks and their position in space. Due to the advantages pointed in Background section, the rig adopted is hybrid. Our rig combines bones with shapes activated by bone's movements.

In this section, we define and describe a novel mapping method to transfer movements tracked by generic MoCap tracking systems to a 3D character and create animation. Our mapping method is composed by mapping parameters and a mapping function. The mapping function is a *runtime* process composed by a Geometric Mapping Algorithm, where we find a correspondence between MoCap's landmarks and the bones in the 3D character's rig, and a Animation Algorithm, where we calculate the intensity of the movement according to rig animation requirements (defined by the artist).

The scheme of the Figure 4.6 provides an overview of our method, showing underlying: inputs, two stages (*calibration* and *runtime*) and algorithms included in each stage. In the *calibration* stage, we calculate the mapping parameters used in *runtime* by the Geometric Mapping Algorithm. Then, to create animation, we explain how the 3D character's rig should be prepared (offline) and how we are able to animate the character using our Animation Algorithm (subsection Creating Animation).

### 4.2.1 Calibration

We start by a short calibration stage to setup the mapping function applied in *runtime*. The *calibration* stage allows the method to collect data from MoCap system and 3D character through the MoCap's and 3D character's parameters, respectively (see Figure

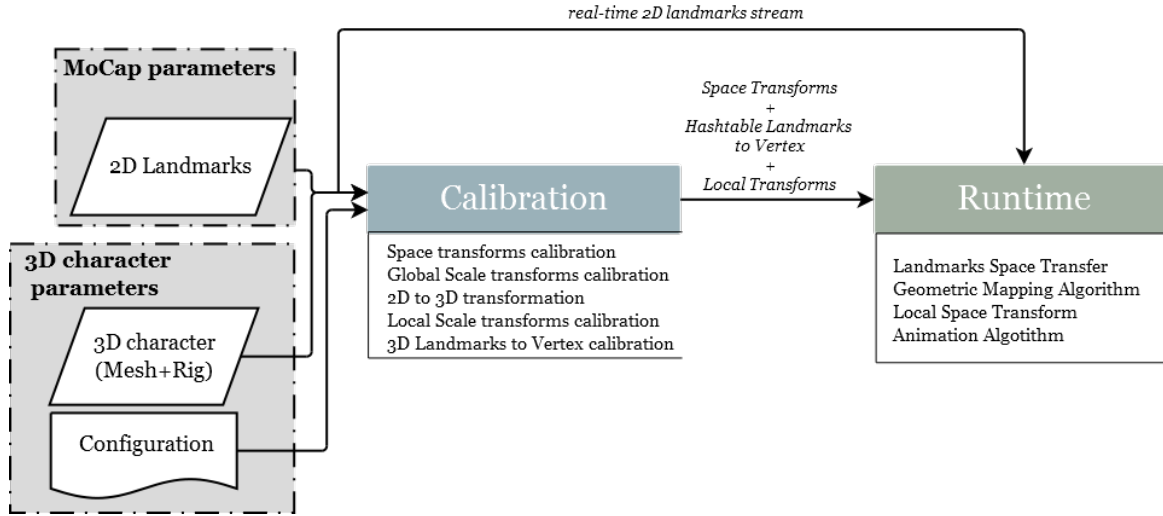


Figure 4.6: Mapping method overview scheme (from left to right): inputs (MoCap and Rig parameters), calibration and runtime stages.

4.7).

MoCap parameters consist in the 2D landmarks with movements captured, while the 3D character's parameters contain both mesh and a configuration file (list of bones with the skin cluster order) generated at subsection Creating Animation during rig setup.

In the first step of *calibration*, we apply a space transformation to make MoCap's 2D landmarks' center match with the tip of the nose of the rig. Afterwards, a global transform (i.e. translation, rotation and scale) is calculated. The global scale is retrieved using the definition of bounding box. The bounding box calculates a ratio between both maximum and minimum values of rightest, leftist, top and bottom values of 2D landmarks and rig's bones (loaded from configuration file). Applying this transformation to 2D landmarks, we "project" them to rig's space. The global transform (i.e. T,R,S Space Calibration in the Figure 4.7) is saved and further applied in *runtime*, with the same goal (i.e. transform MoCap data to match the rig's space).

The second step calculates the third coordinate of 2D landmarks in rig's space. As a solution to get this coordinate, we apply a Raycast method from 2D landmarks to the mesh. The Raycast algorithms give a hit point of the collision, and the triangle where landmarks collided, with respective vertexes [Rot82]. To each one of the raycasted 3D landmarks, we calculate the vertexes' index of hit triangle closest to the hit point, which give us the correspondence between each landmark of MoCap tracker and a vertex of the mesh. Landmarks to Vertex information is saved in a Hashtable. The third coordinate calculated through Raycast is never changed at the *runtime* stage.

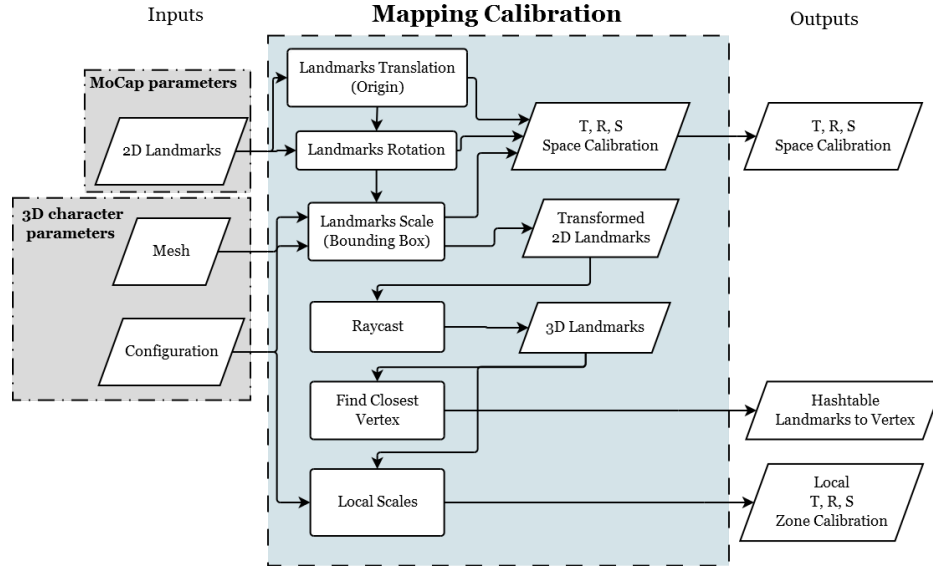


Figure 4.7: Mapping method - calibration stage receives MoCap and 3D character parameters and uses the algorithms at rectangular boxes to return a T,R,S Space Calibration transform (Global and Local) and an Hashtable containing the correspondence between the landmarks and the vertex of the 3D character’s mesh.

Figure 4.8 shows 2D landmark (red) and the 3D landmarks obtained through Raycast (white).

At last, after loading the configuration file containing information about rig’s bones, we ask the user to select the colored points of Figure 4.9 in the 3D character. This is the unique user-dependent calibration of our method and it is only executed one time per 3D character. These positions are used to calculate the local scales in eyebrows and mouth in order to scale locally the movement (i.e. translation, rotation, scale represented as Local T,R,S Zone Calibration in the Figure 4.7) of these facial zones according to users face morphology.

As an output, the calibration stage retrieves the following mapping parameters: T,R,S Space Calibration, Hashtable Landmarks to Vertex and Local T,R,S Zone Calibration. These parameters contain information about MoCap landmarks in the rig space and their connection to the 3D character’s mesh vertexes. We highlight that the *calibration* stage is only executed one time per 3D character.

### 4.2.2 Runtime

This stage transfers the 3D movements captured by the MoCap tracker to the 3D character’s rig, generating animation. Observing Figure 4.10, as an input, this stage

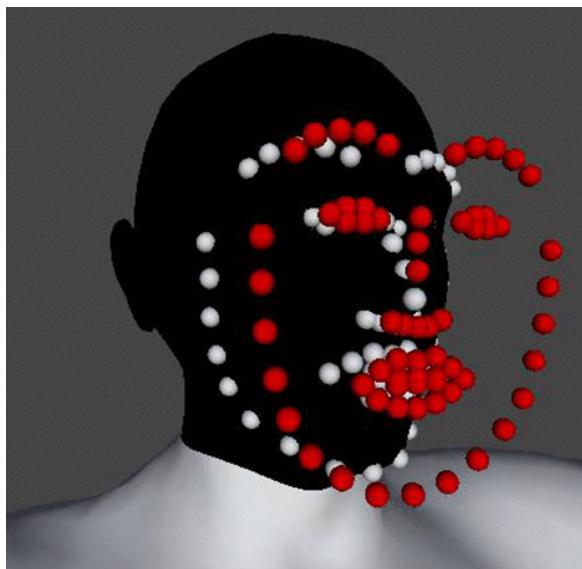


Figure 4.8: Mapping method - Raycast method to calculate the third coordinate of 2D landmarks. Red dots are the input 2D landmarks and the white dots are the resultant 3D landmarks.

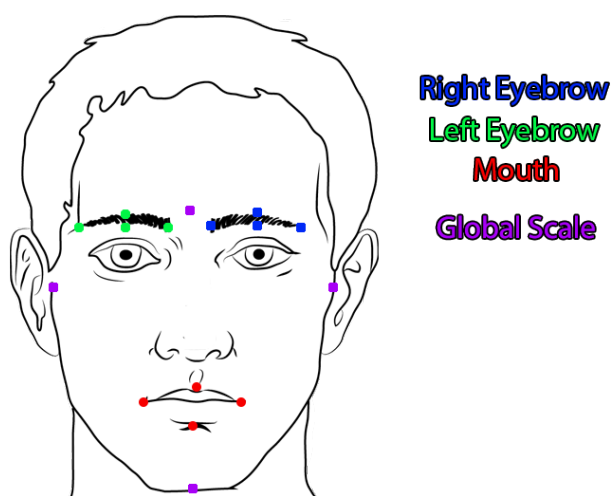


Figure 4.9: Mapping method - Colored landmarks that user selects in the 3D character.

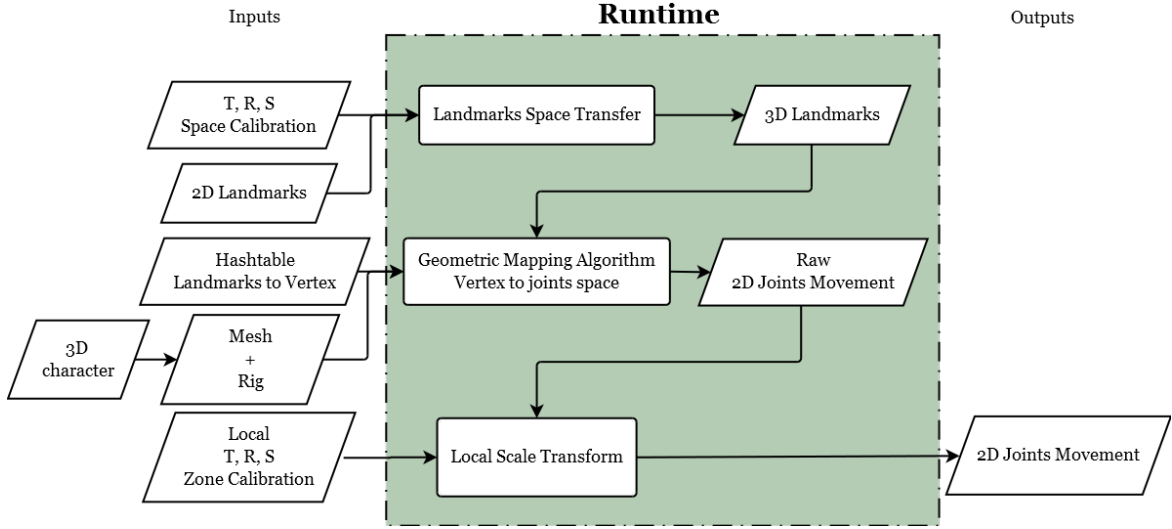


Figure 4.10: Mapping method - Runtime stage that uses the 2D landmarks stream from the MoCap tracking system, the outputs from calibration and 3D character (mesh and rig) to calculate 2D joints movements that are used in the Animation stage.

uses streamed 2D movements and the mapping parameters from *calibration* stage.

Similarly to *calibration*, applying T,R,S Space Calibration transform to 2D landmarks streamed by the MoCap, we obtain the 3D landmarks in rig's space (i.e. Landmarks Space Transfer algorithm). The result of this operation is used as input of the new geometric mapping function. After using the mapping function, the Local T,R,S Zone Calibration is applied to Raw 2D Joints Movement (output from this function) to scale locally the movement, adapting the intensity of the movements to 3D character's topology.

#### 4.2.2.1 Geometric Mapping Algorithm

The mapping function is the core of our novel method and is composed by two algorithms. At first, the Geometric Mapping Algorithm automatizes the process of transference between the 3D landmarks of MoCap tracker to a rig with different bones topology and distribution in space. Secondly, the Animation Algorithm calculates the proper movement intensity according to rig requirements defined by the artist (see subsection Creating Animation). This subsection describes the first algorithm.

The algorithm proposed makes possible the usage of MoCap tracker's landmarks with different structure configuration and number (e.g. if we use only a mouth MoCap, the algorithm finds the connection to rig and animates only the mouth).



At the *calibration* stage, we calculated the correspondence between the landmarks and a vertex in the mesh and the information is stored in a Hashtable Landmarks to Vertex. Now, the goal is to find the correspondence between each vertex and one or more bones in the rig. Note, this algorithm only takes into account the bones that are tagged as animable by the artist (i.e. tagged-bones). For simplicity, we present an example for one vertex, being the process repeated to all the vertexes contained in the Hashtable.

As notation, we define:

- $L_{xyz}$ : Current translation of associated landmark;
- $Li_{xyz}$ : Initial translation of associated landmark;
- $B_{xyz}$ : Current translation of tagged-bone  $B$ ;
- $Bi_{xyz}$ : Initial translation of tagged-bone  $B$ ;

So, to each vertex:

1. We search for the tagged-bone that has more influence in this vertex. This bone is called **main-bone**  $B_M$ . To obtain the translation of  $B_M$ , we calculate:
  - (a) the sum of all weights of  $B_M$  and respective *tagged-bone childs* which also have influence in this very same vertex -  $W_{main}$ ;
  - (b) the translation of  $B_M$  is given by:

$$B_{Mxyz} = \frac{L_{xyz} - Li_{xyz}}{W_{main}} + Bi_{Mxyz}$$

2. We search for all the *secondary-tagged-bones*  $B_{sec}$  that have influence in this vertex and execute, each one:
  - (a) the sum of all weights of  $B_{sec}$  and respective *tagged-bone childs* which also have influence in this very same vertex -  $W_{sec}$ ;
  - (b) the translation of  $B_{sec}$  is given by:

$$B_{sec-xyz} = (B_{Mxyz} - Bi_{Mxyz}) * W_{sec} + Bi_{sec-xyz}$$

Repeating the described process to each vertex (associated to each landmark), we obtain all the tagged bone translations. If a bone has influence from more than one vertex, an average of all the translations assigned from points 1) and 2) is calculated,



giving the final Raw 2D Joints Movement. This step of the method provides the geometrical connection between the MoCap landmarks in rig's space and the bones of the rig, making the mapping method independent of the MoCap system. As already mentioned, the *intensity* of the translation produced and animation generated is explored later at the subsection Creating Animation.

**Note:** the algorithm is applied to calculate the translation of all rig bones except to the jaw bone. To the jaw, we calculate a rotation instead of a translation. Therefore, as a solution, we calculate jaw position through *direct-mapping* (i.e. the direct association between landmark and bone is inserted in configuration file during rig setup. See subsection Creating Animation).

#### 4.2.2.2 Creating Animation

Previous algorithm allows the connection between MoCap tracker and rig. However, to create proper deformation of the 3D character in animation we need to setup the rig (see Rig Setup) and calculate the intensity of the movement produced in each rig's bone in *runtime*.

**Rig setup:** This is a offline process executed one time per rig. The artist starts by providing information about the space where we apply the animation weights, i.e. the artist should define delimited facial areas containing animations (i.e. shapes) (Figure 4.11). We call it animation space. Therefore, the 2D Joints Movements generate animation weights according to the configuration of the respecting animation space (Figure 4.11). Secondly, in our methodology we created a script that is applied to the 3D character and retrieves the Configuration file required for *calibration* stage.

**Animation Algorithm:** 3D character's animation is a *runtime* process (Figure 4.12). Transferring the 2D Joint movements to correspondent animation space, we are able to calculate the proper animation weights responsible by the *intensity* of the movement. The 2D Joint Movements in the animation space are called 2D Movements per Facial Area. After that, we blend the animation created by these joint movements (i.e. Blend Animation), taking into account 3D Animation Structure of the rig (i.e. Animation List and Animation Curves).

More in detail:

- **Weight Animation:** Weight Animation per Facial Area is calculated using the 2D Movement per Facial Area. As already explained, each Facial Area is defined

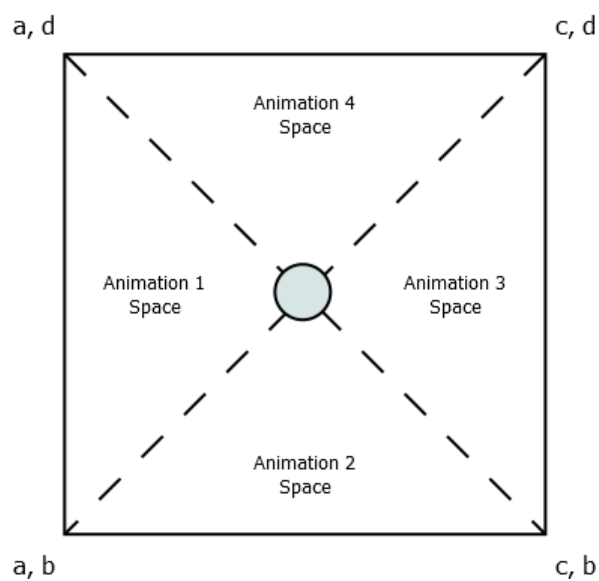


Figure 4.11: Mapping method - Animation Space definition, where the blue dot is a 2D joint.

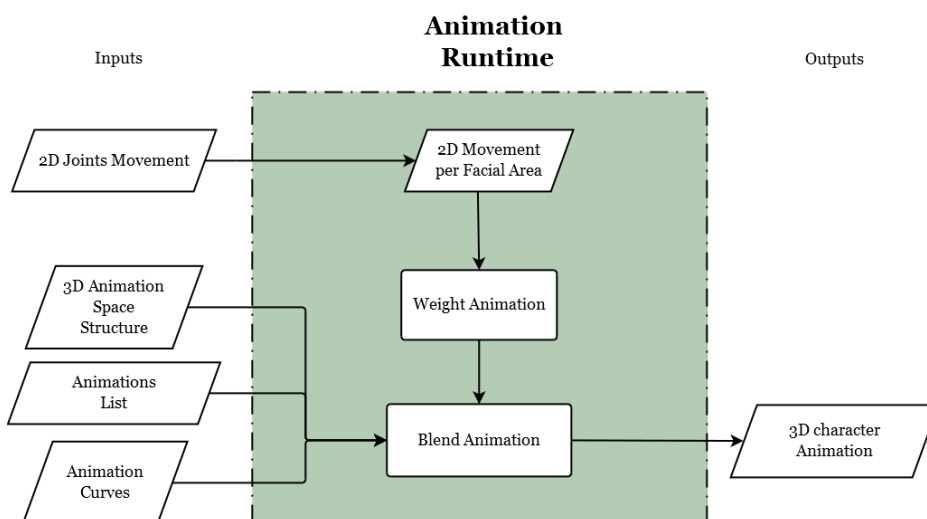


Figure 4.12: Mapping method - Animation Runtime scheme that uses rig information and the 2D joints movements calculated in runtime to produce 3D character's animation.

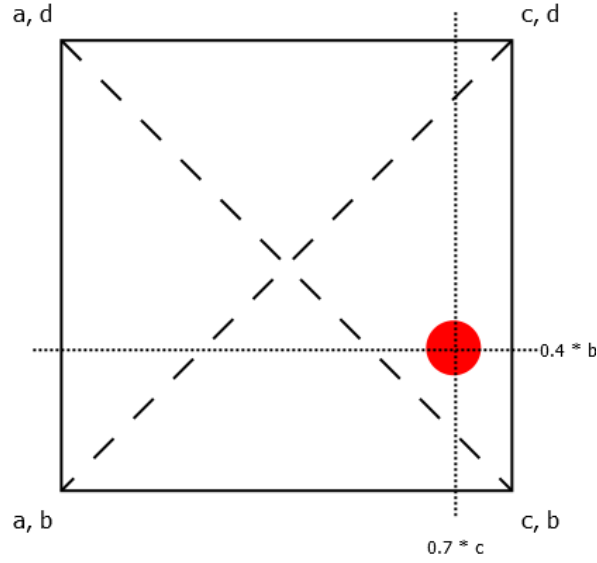


Figure 4.13: Mapping method - Animation weight calculation example, where the red dot 2D translation of the joint of control.

by a rectangle (Figure 4.11). So, the question to be answered is: *How and how much these animations will be activated by a certain movement?* Observe the Figures 4.11 and 4.14 during the further explanation: An Animation Space of certain Face Area is delimited by the points  $(a, b)$ ,  $(c, b)$ ,  $(c, d)$  and  $(a, d)$  (Figure 4.11). Within this rectangle we have 4 animations available, equally distributed in space. The values  $a$ ,  $b$ ,  $c$  and  $d$  delimit the intensity of Animation Space 1, 2, 3 and 4, respectively. Therefore, there is no way to activate the 4 animations simultaneously. An example is represented in Figure 4.14. The red dot is the 2D translation from the joint of control (derived from a certain 2D Joint Movement). In this case, the animation weight  $w_i$  is given by:

$$w_i = 0.4 * b + 0.7 * c$$

In this case, the 2D joint of control position activates Animation 2 with an intensity of  $0.4 * b$  and Animation 3 with an intensity of  $0.7 * c$ .

- **Blend Animation:** Using previous example (Figure 4.14), the next step is to blend Animation 2 and 3 to create a final animation. Here, we use the 3D Animation Structure, i.e. the Animations List and the Animation Curves (Figure shows an example of animation curve). This Structure contains information about animations available and their structure. Using the calculated weights of each animation activated, our methodology make their interpolation and create an unique final animation resultant from the respective facial movement detected.

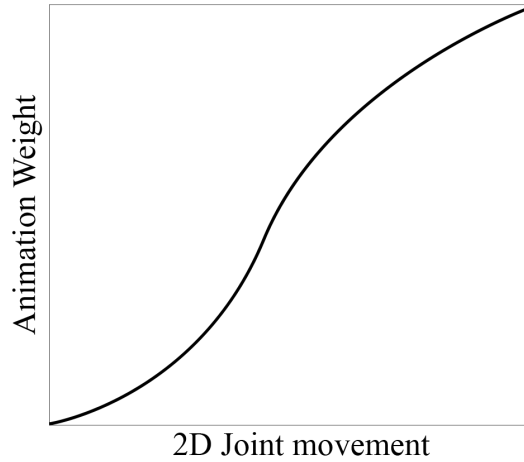


Figure 4.14: Mapping method - Animation Curve example.

**Note:** As explained in the previous section, jaw bone is rotated using *direct-mapping*. Therefore, to create animation the jaw bone orientation follows the associated 2D landmark using a look-at constraint (i.e. the jaw’s rotation is locked so that one of its axes points toward the 2D landmark used as target object.)

### 4.3 Results and Discussion

Using the mapping methodology introduced in this chapter, we built an application in Unity 3D engine ©. The application allows the inclusion of different facial MoCap systems and 3D characters. As proof-of-concept and validation, we tested our mapping method using (i) our facial MoCap tracker (Chapter 3) and (ii) an offline MoCap tracker using a HMC video.

Figure 4.15, 4.16, 4.17 and 4.18 show examples of the animations obtained using real-time facial MoCap (using our approach introduced at the chapter 3) and our novel mapping method.

Overall analysis show that 3D character reproduced the movements captured by the MoCap tracker, even if they are uncommon facial traits like asymmetric movements (see Figures 4.16 and 4.17). Thus, the mapping method combined with our hybrid rig approach leads to realistic behaviors, even when peculiar movements occur.

However, in Figure 4.18, we observe an undesirable behavior of inner lips during a surprise-like expression. The undesirable result comes from the mapping method

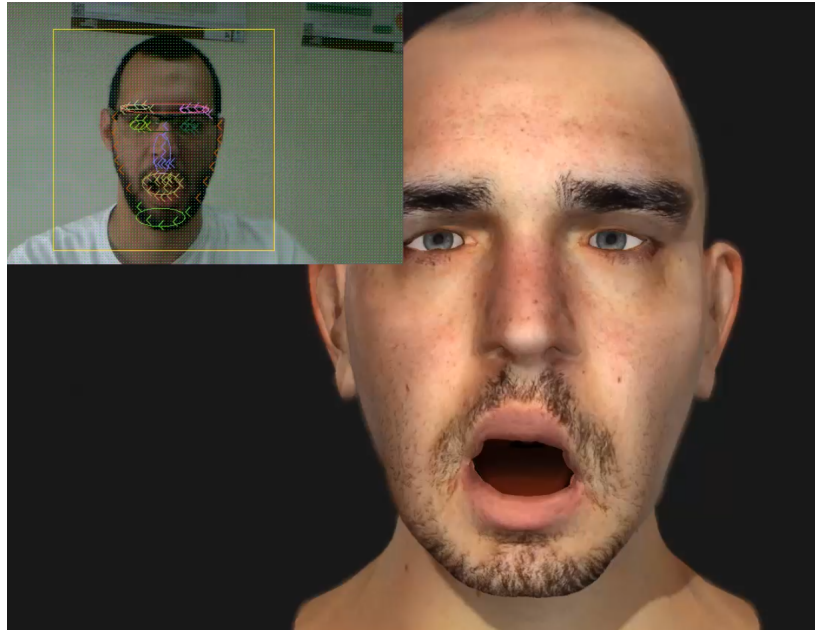


Figure 4.15: Mapping results - Surprise facial expression reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time.

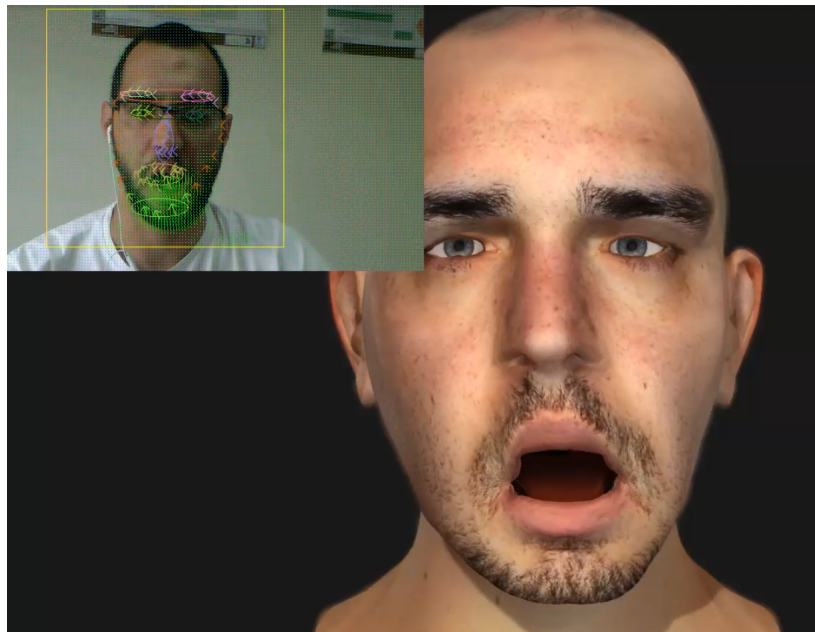


Figure 4.16: Mapping results - Surprise facial expression with jaw asymmetrical movement reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time.

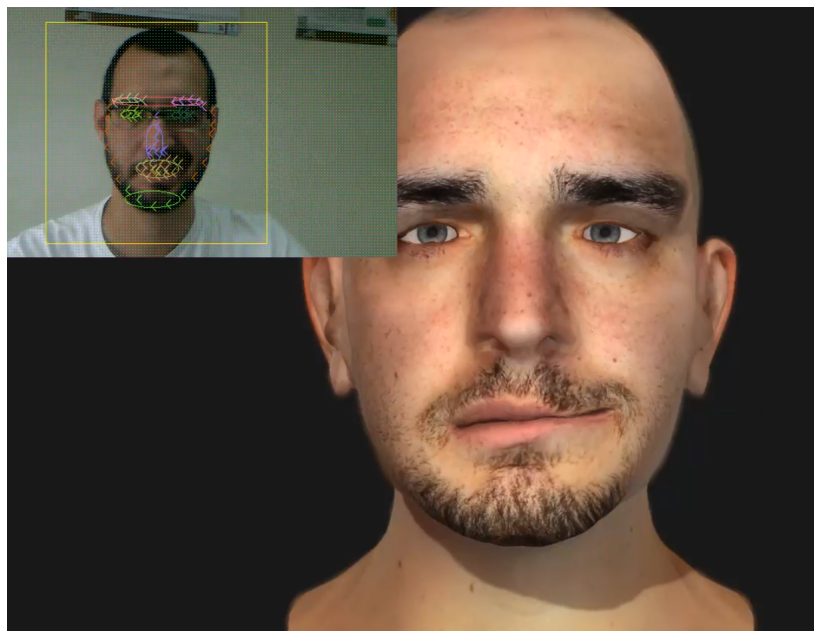


Figure 4.17: Mapping results - Surprise facial expression with jaw asymmetrical movement reproduction in 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time.

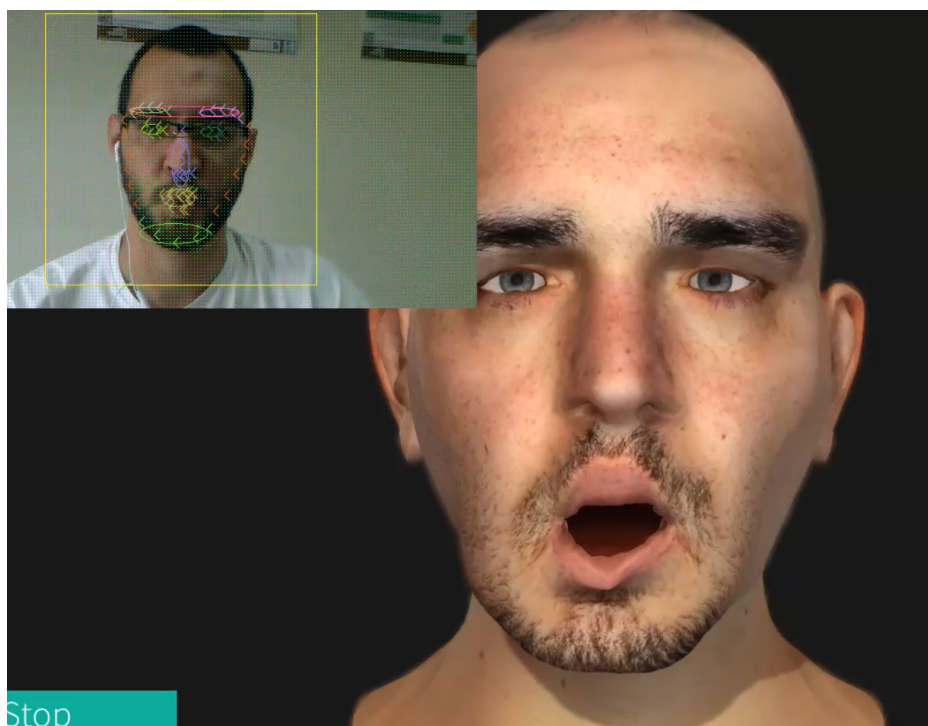


Figure 4.18: Mapping results - Example of inner lip undesirable effect in the 3D character (right) using our facial MoCap approach (left). Capture and mapping in real-time.

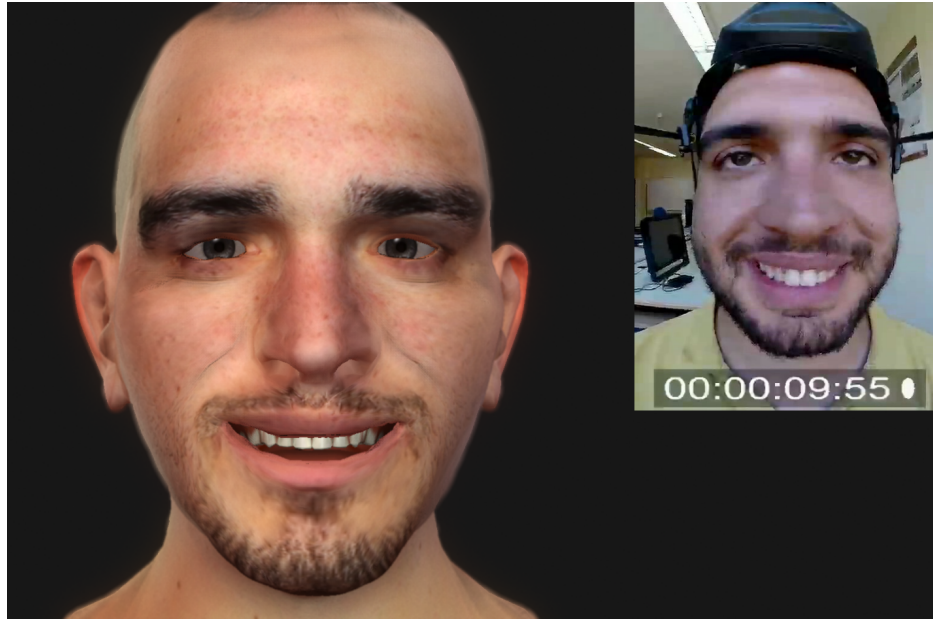


Figure 4.19: Mapping results - Example of joy-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.

incapacity of removing a bad tracking behavior of the MoCap tracker. Therefore, some refinements should be done in the mapping function to avoid their high sensitivity to tracker errors. The undesirable deformation arise when adjacent bones were activated too differently, resulting in an "edge" in the mesh. So, this error arise from the second part of the mapping function, in the Animation Algorithm. A possible solution is to apply a smooth algorithm after weights calculation in animation, like Kalman or Laplacian smoothing or more advanced smoothing algorithms [SZG15], to reduce these discrepancies in deformation of adjacent vertex. To test the mapping method itself, we deployed a test using an high accuracy facial MoCap applied to a HMC video (i.e. Nuke ©Offline Optical Flow ). The adopted facial MoCap requires the manual placement of landmarks and the offline tracking of their positions. As output, the tracker delivers accurate 2D landmarks' positions to each frame of the video. The frame landmarks were used in real-time as input in our mapping method. Results are in the Figures 4.19, 4.20, 4.21, 4.22 and 4.23.

The results for real-time mapping using the offline tracker show that our mapping method allowed proper reproduction of movements captured in the 3D character (see Figures 4.19, 4.20, 4.21 and 4.22). We observed that our method has a smoother behavior using an accurate MoCap tracker. This confirms previous statement: our mapping method still presents sensitivity to facial MoCap tracker's performance.

To test the efficiency plus accuracy of the method reproducing subtle and fast move-



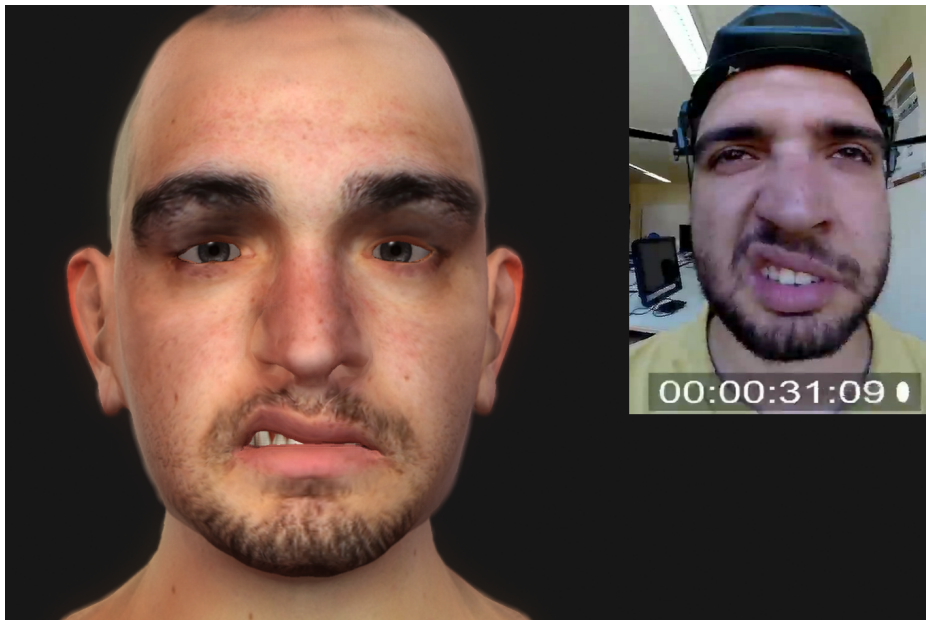


Figure 4.20: Mapping results - Example of disgust-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.

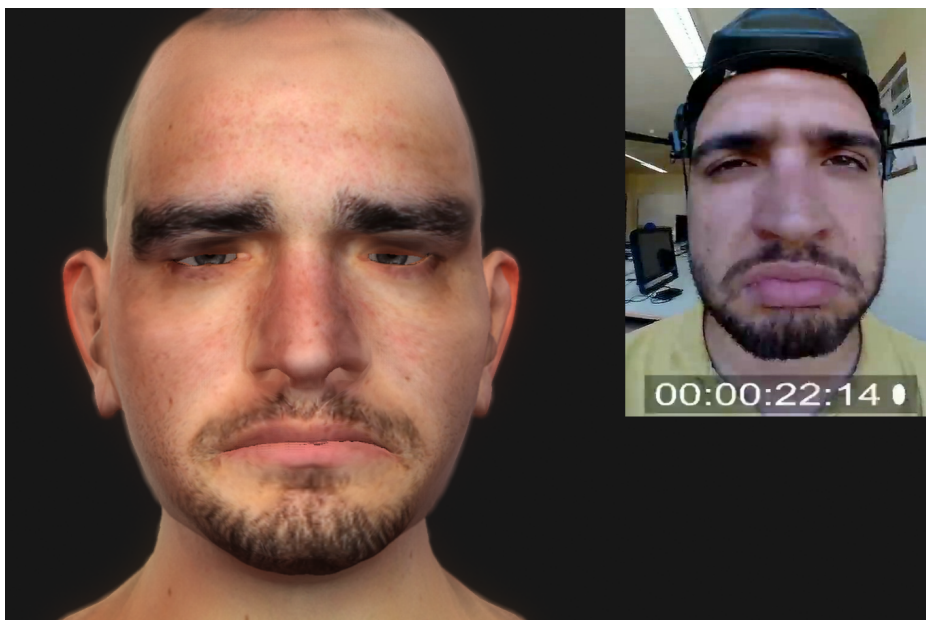


Figure 4.21: Mapping results - Example of sadness-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.



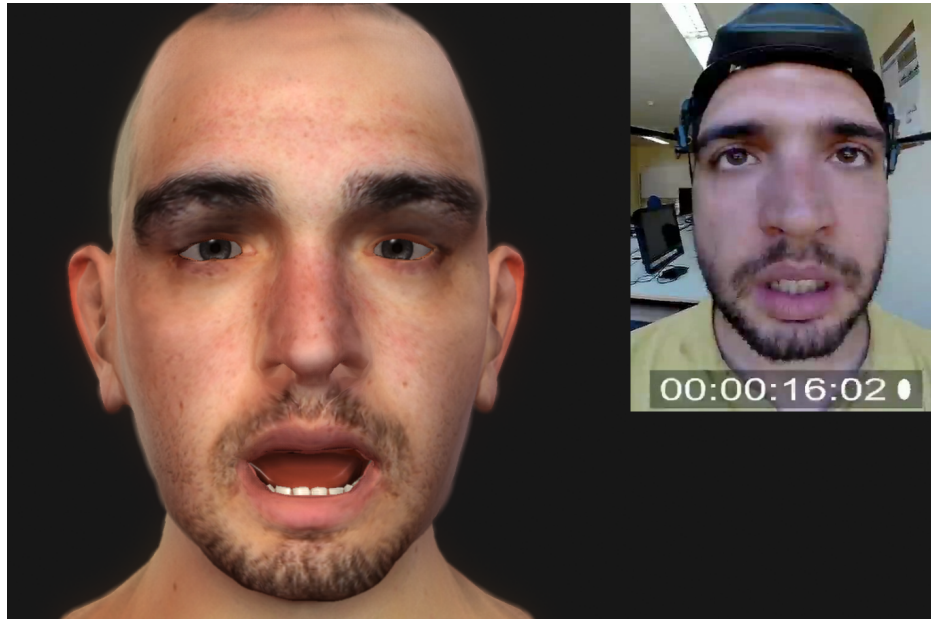


Figure 4.22: Mapping results - Example of surprise-like expression animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.

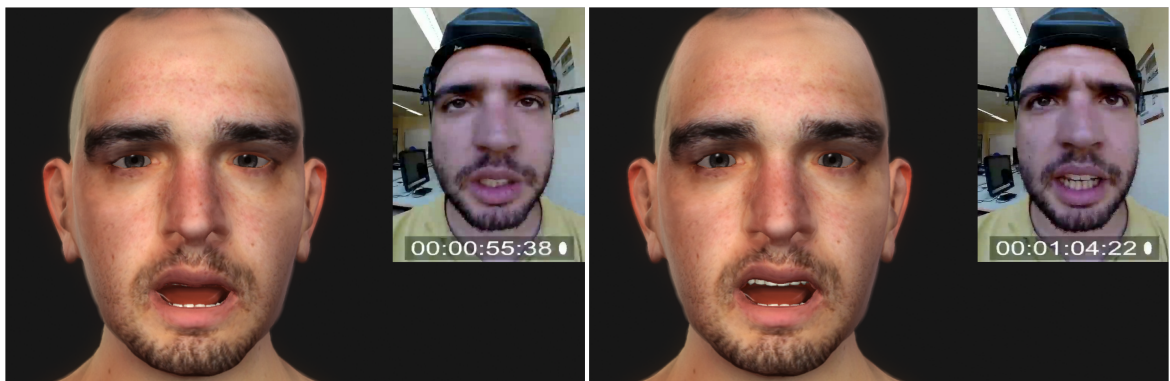


Figure 4.23: Mapping results - Example of speech sequence animation (left) using offline facial MoCap (right). Offline capture and mapping in real-time.

ments, we recorded a speech video and applied the offline MoCap. In Figure 4.23, we show a sequence of two speech shots. The mapping method’s latency does not interfere in animation. The method is able to generate animations fast enough, however, some subtle movements are lost. Looking into the eyebrows positioning in video and the 3D character’s face, we observe the ”loss” of movement. This behavior is due to the calculus of translations average when a bone influence more than one vertex. In this case, the suggested smoothing may erase even more the small movements. Therefore, as future work, we suggest the deployment of optimizations when a bone influences more than one vertex (e.g. use median calculus) combined with a local smoothing algorithm. The goal is to reach a balance where we are able to reproduce subtle movements and, simultaneously, eliminate undesirable movements.

## 4.4 Conclusions

This chapter describes a novel methodology for MoCap facial animation (i.e. mapping method). As major benefits, we retrieve a modular method: (i) adaptable to different facial MoCap trackers and (ii) allows real-time facial animation without complex calibrations. Our method executes a calibration stage one time per 3D character, where the user-dependent step is related to the placement of 16 markers in the 3D character. Regarding movements transfer, our MoCap facial animation method enables the real-time mapping between user’s facial movements and the 3D character, even when asymmetrical movements occur (e.g. mouth moving to the left). Yet, our method presents difficulties in the animation of subtle and fast movements (i.e. speech) . As a possible solution, we suggest the refactoring and parallelization of the mapping method. Through the increase of computational performance of the mapping method, the 3D character will be able to reproduce the user’s facial movements in a faster and more responsive way.

# Chapter 5

## MoCap VR Methods

*"The computer is a protean technology; virtual reality is a protean medium. As virtual environments begin to diffuse throughout society, the range of these systems will be quite broad. The categories proposed here will certainly increase in complexity. The categories of components and the distinctions among systems will multiply as the virtual environment marketplace bursts into a kaleidoscope of applications and options. Like the microchip, a version of this medium may find its way into almost every form of mediated communication. From the low-end to the high-end system, various configurations of these components may grow to simulate every communication channel from a handshake, to a book, to the video image."*

Frank Biocca and Mark R. Levy  
Communication in the Age of Virtual Reality (1995)

*This book was written in 1995. By that date the authors were already aware of the impact of Virtual Reality (VR) technologies in human communication. However, HMDs only started to be distributed to mass-market in 2014, by the company Oculus VR. The revival of VR created new needs and challenges in HCI, CV and CG field, that needed to adapt and provide compatible digital media and interaction tools for a complete sense of embodiment. As already explored in this PhD thesis, facial expressions are the main non-verbal communication channel used by humans. This lead to an urgent demand of new facial recognition and capture approaches to extract facial expressions that are compatible with occlusions created by VR headsets. The new approaches will activate on-the-fly 3D character animation improving face-to-face communication and interactions inside VR environments. In this chapter, we propose and validate a novel methodology for automatic facial expressions estimation*

*compatible with VR scenarios.*

---

**Reference:** This article was published at the VISAPP 2016 11st International Conference on Computer Vision Theory and Applications (Appendix F).

## 5.1 Background

Today, we live a continuous evolution of global digital interactions and communication between humans [JJ13]. There are no geographic barriers. We communicate using phones, computers and, more recently, we are able to communicate inside VR environments using VR headsets (i.e. HMD) (Oculus VR 2014). Using phones and computers, we are able to transmit real-time speech and video stream, having a both communication channels (i.e. verbal and non-verbal). Although, to achieve this level of communication and interaction in VR environments, we need to have: (i) a virtual avatar (i.e. a 3D character) and (ii) to develop advanced capture and animation systems to animate and control these characters in real-time [Bio97, Sla14]. These on-the-fly interaction using 3D characters enhances the three components of the sense of embodiment: self-location, agency and body ownership [Bio97, KGS12]. In the previous chapters, we built methods to automatize 3D characters' MoCap animation. However, these methods are not compatible with VR communication scenarios due to the occlusions produced by the VR headsets. Consequently, this section aims to explore the occlusion problem and create a compatible methodology to access facial expressions in both occluded and non-occluded face's regions. Facial expressions recognized can be further applied to trigger animation in 3D characters or be used as interface for emotion-based tools, like emotional gaming (e.g. Left 4 Dead 2 by Valve).

As Background to guide our method deployment, we study the literature regarding two different topics: (i) facial MoCap solutions for persistent partial occlusions created by VR Head Mounted Displays (HMD) and (ii) partial occlusions impact in facial expressiveness. The first topic presents state of the art facial MoCap solutions to overcome the persistent occlusions' issue. Then, in (ii), we show how these occlusions restrict face-to face communication and how they impact the face expressiveness, searching for a connection between facial movements in occluded and non-occluded regions.

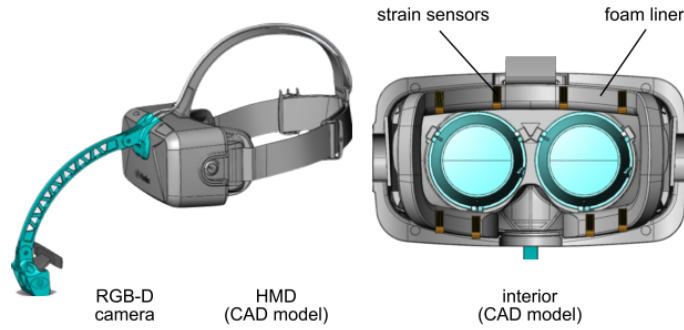


Figure 5.1: VR hardware-based setup proposed by Li *et al.* [LTO<sup>+</sup>15] to overcome partial occlusions issue.

### 5.1.1 Persistent Partial Occlusions: a today's problem

As stated in the previous chapters, in literature there are several promising solutions for real-time automatic facial tracking. Although, with the arise of VR commercial approaches of consumer-level HMD's (Oculus VR 2014), a new issue was raised: the real-time automatic tracking of faces partially occluded by hardware (i.e. persistent partial occlusions of face) [Sla14]. Current MoCap approaches do not support persistent partial occlusions presenting error accumulation [CHZ14]. Therefore, due to the absence of VR devices in mass-market, the occlusion issue remained unsolved. Only in 2015, Hao Li *et al.* [LTO<sup>+</sup>15] highlighted this problem and proposed a hardware based tracking solution. Li *et al.* [LTO<sup>+</sup>15] uses an RGB-D camera combined with eight ultra-thin strain gauges (flexible metal foil sensors) placed on the foam liner for surface strain measurements to track upper face movements, occluded by the HMD (see Figure 5.1). As first limitation, this approach relies on a long initial calibration to fit the measures to each individual's faces using a training sequence of FACS [EF78]. Also, in subsequent wearings by the same person, a smaller calibration is needed, to re-adapt the hardware measures. These training steps allow the detection of user's upper and bottom face expressions and activate a blendshape rig containing the full range of FACS shapes [EF78]. Besides the complexity of usage, Li *et al.*'s [LTO<sup>+</sup>15] work pinpointed drifts and accuracy decrease due to variations in pressure distribution from HMD placement and head orientation. As consequence, the HMD head positioning influences eyebrows' movement detection.

### 5.1.2 Partial Occlusions and Expressiveness

Humans' communication rely in facial expressions and emotions to transmit and enhance information not provided by speech [LWHW12]. Even communicating through technology, we always search for a way to use the non-verbal communication channel. As example, using video stream of our faces or virtual representations, like cartoons or 3D characters with pre-defined facial expressions, etc. Understanding facial expressions and their believable reproduction in 3D characters is one of the key challenges of CG and plays an important role in digital economy [JJ13]. This role is even more relevant now, with recent advances in VR communications at consumer level [Bio97]. *But how can we use the common solutions of facial animations, like MoCap, if user's face is occluded? Are we able to represent faces using information only from bottom of the face?* To answer these questions, we make a literature overview regarding face regions impact in non-verbal communication. The goal is to understand how a partial occlusion of the face affects communication. We also researched for a relationship between occluded and non-occluded facial parts through biomechanics and emotion-based studies. This information was used to build this chapter hypothesis.

In Chapter 2 - MoCap Fundamental Science, we deliver two studies to explore face role in human communication and expression of emotions. During those studies, we used the face as a whole, without restriction to certain regions. Hence, our face perception study showed an independent shape representation of upper and bottom parts of the face. Similar conclusions are found in emotion perception's literature, where mouth and eyes play different roles [EA11, LWHW12, BSSM<sup>+</sup>13]. In [EA11, BSSM<sup>+</sup>13] it is shown that according to the emotion detected participants used information from eyes, or mouth or both. This is, in happy expressions participants used information from the mouth; for sad and angry, from eyes; and to fear and neutral, both mouth and eyes are used. For additional information about non-verbal communication, the reader can access [LWHW12]. Taking these statements into account, if we occlude certain facial region the communication is affected and we may not be able to decode expressions properly. Subsequently, the tracking of only certain part of the face is not enough for a proper communication and to generate believable facial animation of 3D characters.

From the biomechanical point of view, facial muscles work synergistically to create expressions. They interweave with one another, being difficult to decode their boundaries, since their terminal ends are interlaced with other muscles. A detailed research about facial anatomy and biomechanics can be accessed in the Chapter 3

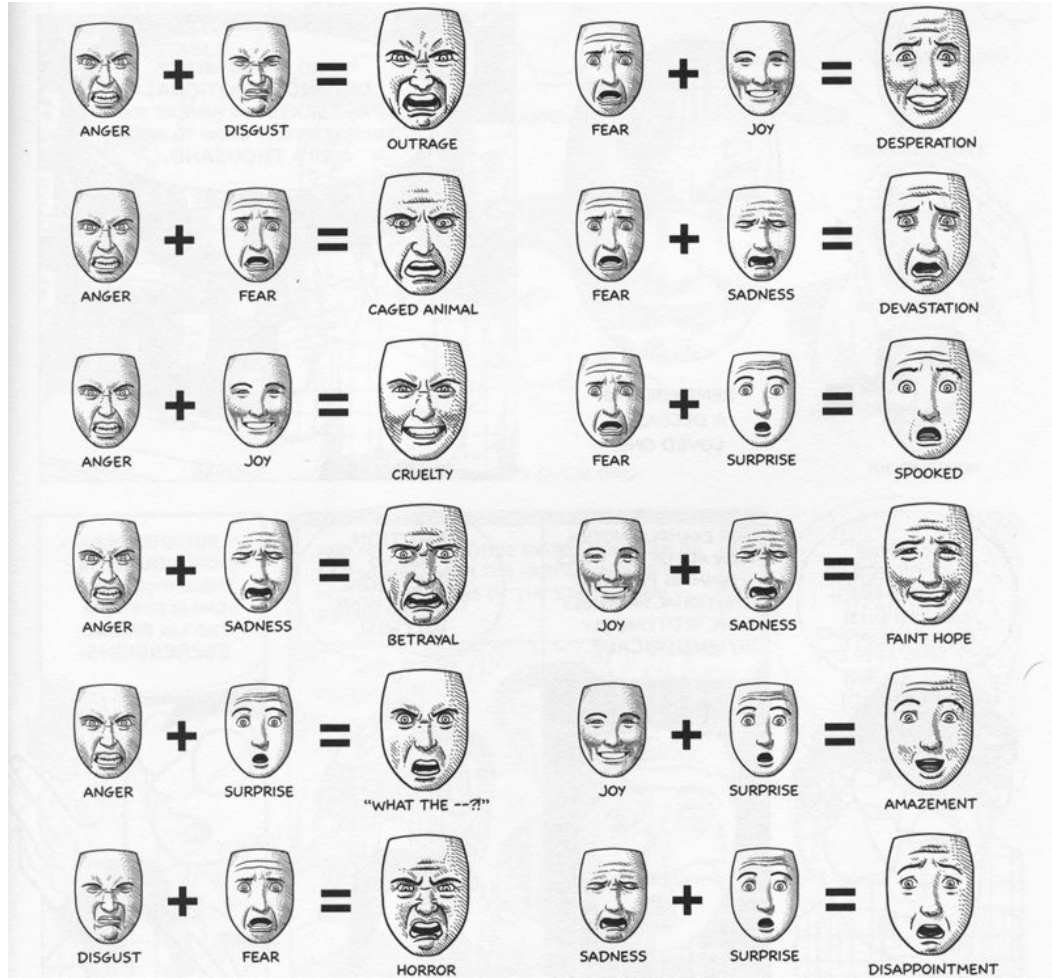


Figure 5.2: Examples of diversity of facial expressions created by mixing two basic emotions [McC06].

of the book *Computer Facial Animation* [PWA96]. Several studies in CG use the biomechanical approach to define coding systems. These coding systems parameterize human face enabling a faster generation of facial expressions in 3D characters [EF78, PF03, MTPT88]. Although, they do not provide a clear solution for facial expressions estimation constrained to certain regions of the face. Furthermore, the definition and prediction of facial expressions is even more challenging when we observe their diversity across individuals. Scott McCloud [McC06] explains the infinite possibilities of facial expressions combinations (i.e. the way mixing any two of universal emotions can generate a third expression, which, in many cases, is also distinct and recognizable enough to earn its own name) [McC06] (see Figure 5.2).

By literature analysis, we attain that occlusions generated by VR devices affect communication and if we limit the tracking of facial expressions to non-occluded regions we

are not able to produce believable animation to cover the diversity of faces. However, biomechanics and facial animation coding systems show a connection between the different facial regions. In the next section, we use the aforementioned statements and describe a novel methodology to overcome occlusions problem of facial MoCap and then, to assess facial expressions using non-occluded face information.

## 5.2 Methodology

Combining the knowledge in facial expressions (perception, recognition and capture) and animation acquired during this PhD research, with previous section final statements, we formulate the following hypothesis:

*to create methods to estimate facial expressions of upper part of the face and predicts emotions using movements tracked from bottom of the face.*

As main goal, we aim to deliver MoCap VR methods, that:

- overcome the persistent partial occlusions issue in MoCap;
- recognize universal emotions, plus neutral [EF75, JJ13], using bottom face features tracking;
- estimate upper face movements (i.e. eyebrows movements) using information from bottom part of the face.

The Figure 5.3 shows the connection between the three MoCap VR methods. Check on Figure 1.3 in Chapter 1 - Introduction, to see how the work delivered in this chapter fits the overall thesis framework.

We start by presenting a method to make generic MoCap systems compatible with persistent partial occlusions produced by VR headsets. Then, applying this algorithm, we are able to track properly the bottom face's features and use them to develop methods that predict the following facial expressions: (i) universal emotions, plus neutral [EF75, JJ13] and (ii) eyebrows movements. Combining aforementioned methods, we make possible the MoCap of upper and bottom face movements and estimation of facial emotions under persistent partial occlusions created by VR headsets.



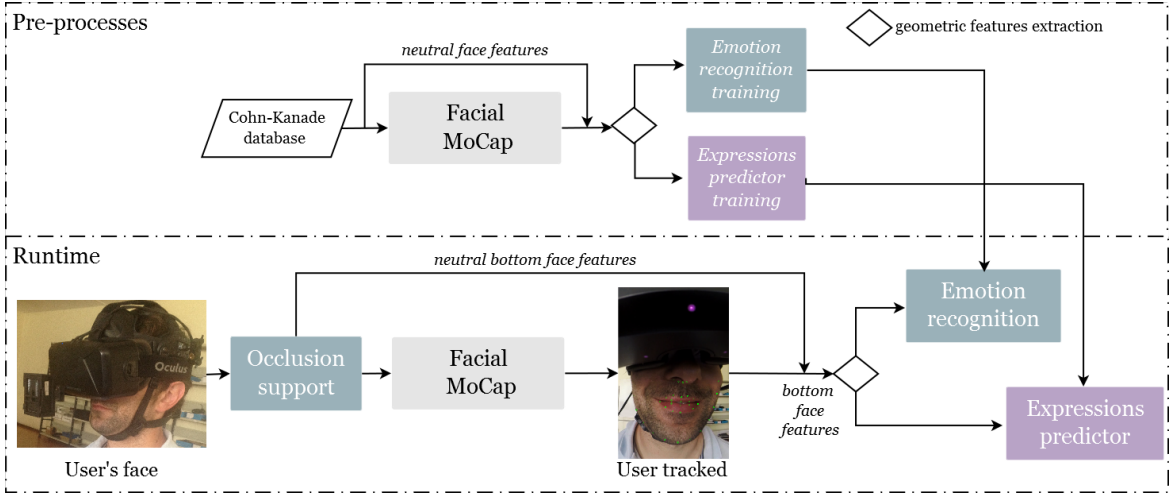


Figure 5.3: MoCap VR methods' framework: filled blue and purple boxes refer to our VR methods distributed as pre-processes or runtime processes.

As setup, we suggest the usage of a Head Mounted Camera (HMC) combined with the VR HMD (see Figure 5.4). At first, we justify the adoption of HMC as capture hardware: When the user is inside the VR environment he is not aware of the space around him. The VR devices precisely substitute the user's sensory input and transform the meaning of their motor outputs with reference to an exactly knowable alternate reality [Sla14]. Hence, the user moves and reacts to impulses from VR environment. If we want to capture his face, we have to attach a capture device (i.e. camera) to his body and the device should follow user's movements (see HMC on Figure 5.4). It is not possible to use a static camera, because the user is not going to be able to place himself in a position proper for capture. A similar setup was also proposed by Li *et al.* [LTO<sup>+</sup>15], but we removed the strain sensors.

### 5.2.1 VR Persistent Partial Occlusions: a novel method

We start by describing a modular method to solve the issue of persistent partial occlusions of the face produced by VR headsets. To deploy our occlusion support method for facial MoCap, we used the following statement: we know the kind of occlusion created by HMD, so we know which part of the face is occluded. We also know that MoCap algorithms fail in these situations because they use a face model. When the face is occluded this model starts not to fit since there is not a full face being captured. As a solution, we use the knowledge that the region occluded is the upper part of the face to "re-create" the whole face.

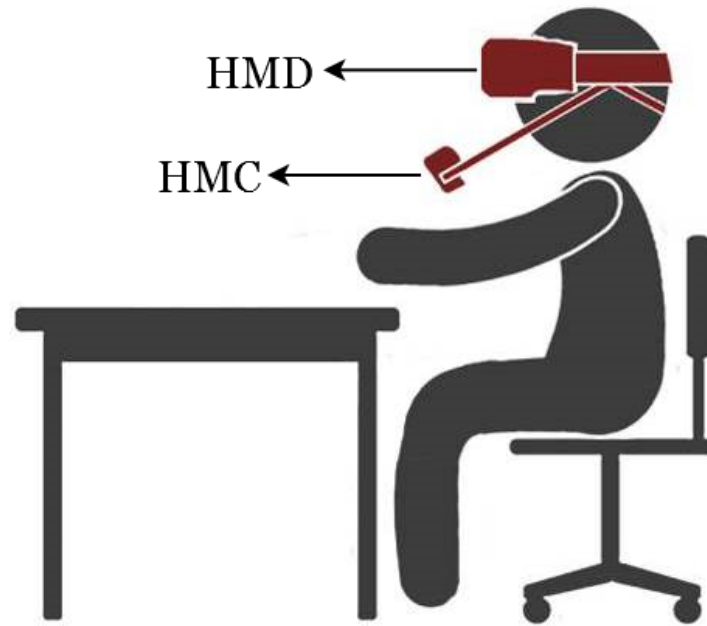


Figure 5.4: MoCap VR hardware setup.



Figure 5.5: VR setup examples with: nVisor SX111 (left) and Oculus Rift DK2(right).

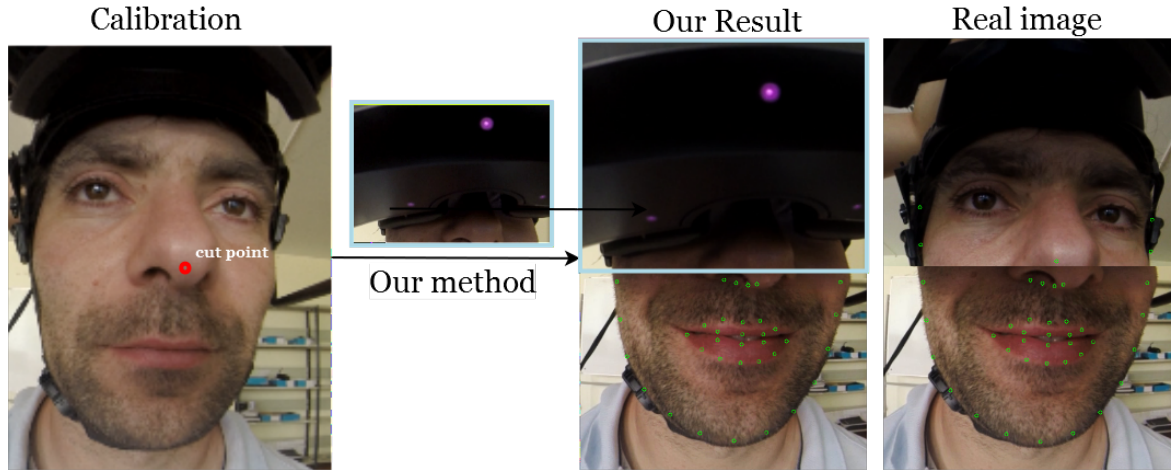


Figure 5.6: MoCap VR method: Persistent partial occlusions. From left to right: calibration image without VR HMD; our method uses cut point (red circle) to cut image an overlay at subsequent images: at left, what facial MoCap method see is a full face and, at right, the real image.

Our novel method overlays the upper part of the face captured on a neutral pose during calibration. Firstly, we assume that the higher visible point of the face is the nose and define it as cut point (i.e. this point can be changed to fit the occlusion created by certain HMD). Then, we detect the cut point with the MoCap and we cut the upper part of the calibration image (i.e. frame streamed) from the nose up, and use it to overlay to all the next camera/video frames. Hence, now the occluded part of the face is replaced with a static neutral face. The MoCap system is now able to detect the features in the combined half static/ half expressive face (see Figure 5.6). We ensure a proper re-creation of a face since we use a HMC that removes the user's head movements, i.e. user's face is in the same position during calibration and next streamed images.

### 5.2.2 VR Assessing Facial Expressions

During the development of the VR facial expressions method, we applied face features and machine learning know-how acquired during the emotion recognition study (Chapter 2 - MoCap Fundamental Science). We followed a similar procedure. The goals here are quite different: real-time emotion recognition of universal emotions [EF75] and upper face (i.e. occluded region) expressions prediction in a VR scenario, where we are only able to track bottom face features. We aim to track facial expressions ahead of only emotions, in order to get a wide change of facial expressions and better cover and representation of the diversity of faces [McC93]. In opposition to the

emotion classification method of Chapter 2, where we needed to reduce the number of features tracked, in VR scenarios we have to maximize the information tracked in the bottom part of the face. Therefore, the feature extraction method should be able to retrieve enough information to allow an accurate prediction of facial expressions by the machine learning algorithm. As a solution, we propose to use all the features tracked of bottom face region (see Figure 5.7 blue rectangle) and apply a geometrical features extraction algorithm. This algorithm is defined as the Euclidean distance between neutral face features (stored during calibration step of previous persistent partial occlusions method) and current frame (i.e. instant in time) features. Summarizing, to each feature tracked  $p$  in certain instant  $i$ , we calculate the distance  $D(p_i, p_c)$ :

$$D(p_i, p_c) = \sqrt{\frac{((p_i(x) - p_c(x))^2 + (p_i(y) - p_c(y))^2)}{\|p_i - p_c\|}}$$

,where:

$p_i$  is the 2D bottom face feature  $p$  at the instant  $i$  in time;

$p_c$  is the 2D bottom face feature  $p$  of neutral expression captured during calibration;

$\|p_i - p_c\|$  is the norm between  $p_i$  and  $p_c$  in Cartesian space.

Since the occlusion produced varies according to VR headset used, we also created machine learning models to assess facial expressions using the bottom face features information including and excluding nose features. The bottom face features without nose feature can be used by the different kinds of HMD, since the nose region is the one affected by the device size.

To create the machine learning models to predict the emotions and upper face expressions, we used the Cohn-Kanade (CK+) database [LCK<sup>+</sup>10]. This database contains posed and spontaneous sequences from 210 participants (i.e. cross-cultural adults of both genres). Each sequence starts with a neutral expression and proceeds to a peak expression. This sequences are FACS coded and emotion labeled. The transition between neutral and a peak expression allowed us to detect spontaneous expressions and not only pure full expressions like in the emotion recognition study in Chapter 2.

Taking into account the results obtained machine learning classifiers comparison of emotion recognition study (Chapter 2), we adopted a GPU version of Random Forest

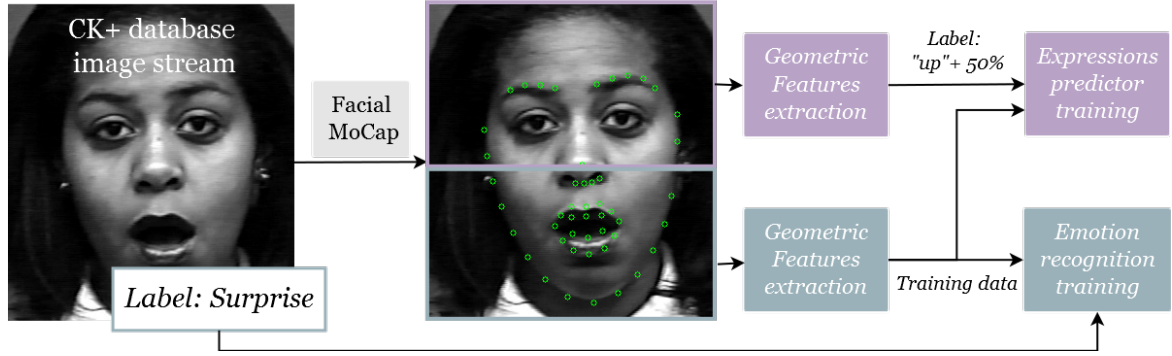


Figure 5.7: MoCap VR methods: Expressions predictor training (purple) and emotion predictor training (blue) with CK+ database.

[Bre01] to create the machine learning models for real-time prediction. To implement these VR methods we used an upgraded version of OpenCV [ope14] in C++ language which included a GPU accelerated version of the machine learning classifier. We also adopted Saragih *et al.* [SLC11a] as MoCap system. (see Figure 5.7 tracking landmarks in green). Note, we did not adopt our facial MoCap method because it does not use a face model and our occlusion support method (previous section) will not be able to reduce completely the error accumulation. Since the main scope in this chapter is to test our hypothesis, we adopted a model-based approach to ensure higher detection stability and whose facial features impact in facial expressions using Random Forests classification was already proven (see study two in Chapter 2).

### 5.2.2.1 VR Emotion Recognition: novel method

In the pre-process stage, we create the Random Forests model that is used to predict emotions in real-time. To build the model for emotion classification, to each database's sequence we applied the facial MoCap method and extracted bottom face features. Using the first frame of the sequence as neutral expression, to subsequent frames in the sequence, we calculate the distance  $D(p_i, p_c)$ , between bottom face features of current frame and neutral expression's frame. Thus, to train the machine learning model for emotion recognition we used aforementioned geometrical extraction algorithm: distance  $D(p_i, p_c)$  of bottom face's features of each frame. As response value, to each distance calculated, we used respective CK+ emotion label (see Figure 5.7 blue processes).

As observed in the Figure 5.4, in the first frame, we apply our occlusions support method and store neutral face features. This step is only execute one time per user.

After, in runtime, the adapted facial MoCap system delivers bottom face's movements and distance  $D(p_i, p_c)$  is calculated to each feature  $p$ . The group of distances are used as input in the Random Forests classifier that predicts the user's emotion represented by that distances and respective accuracy's percentage.

### 5.2.2.2 VR Facial Expressions Predictor: novel method

To train the upper face expressions model, we also used the distance of neutral and expression bottom face features as geometric extraction algorithm. However, we have to define the movements that we wanted to predict in order to create specific tags to the training process. For simplicity, we set as upper face expressions the prediction of eyebrows movements, i.e. the detection if eyebrows are going up or down, and the how much they are moving compared to a neutral position. This last parameter is measured as a percentage of movement up/down compared to neutral expression. Similarly to the assumption made at face self-perception study of Chapter 2, we assume symmetry of the eyebrows movements [FRB<sup>+</sup>13]. To define the tags, we calculated the Euclidean distance  $D(p_i, p_c)$  between neutral position of eyebrows and the expression positions in the other frames of the sequence. If the average of the eyebrows features indicated that they are going up, we tagged "up"; the opposite if the eyebrows went down we tag "down" (i.e. we used image coordinate system, so this distance was negative when eyebrows go up and vice-versa). Simultaneously to each frame of the sequence tagged we saved the percentage of movement compared to neutral position (up or down). As result, to each frame of the sequence of each participant in CK+ database, we tagged: eyebrows "up" or "down", plus percentage of movement. In the Figure 5.7 with purple processes, the reader can observe an example of method's framework.

At pre-process stage, we trained two Random Forests models with the same input data: the distances  $D(p_i, p_c)$  between neutral and current bottom face features; but using one of the following response values:

- "up" and percentage of movement, if eyebrows are rising
- "down" and percentage of movement, if eyebrows are descending

, to each frame of each sequence of CK+ database.

Since we are using a GPU approach of the classifier with high computational performance, to reach higher levels of accuracy in eyebrows movements' prediction we trained two different models: one to predict the rise movement and, another to predict the

opposite. Validation and results are displayed in the next section (i.e. Results and Validation).

In runtime, we apply the defined geometrical features extraction to the bottom face's features tracked by the adapted MoCap. The extracted features are used as input in both Random Forests classifiers that retrieve, in real-time, one of the predictions:

1. **eyebrows "rising"** and percentage of movement;
2. **eyebrows "descending"** and percentage of movement.

Since we are using two different classifiers, there is a probability of confusion of both models return simultaneously an "up" and "down" movement. As a solution, our method compares the accuracies of prediction from the two classifiers' predictions, and the result delivered is the one with higher accuracy.

## 5.3 Results and Validation

In this section, we show the results and statistical validation of the methods proposed. Statistical analysis was performed using R software [?].

### 5.3.1 MoCap VR method: Persistent Partial Occlusions

To test our occlusions method, we applied it to Saragih *et al.* [SLC11a] and Cao *et al.* [CHZ14] MoCap systems (see Figures 5.8 and 5.9, respectively). At the Figure 5.10, we test a generic partial occlusion created by a piece of paper.

As observed in the Figures 5.8, 5.9 and 5.10, our occlusion-support method adapts to MoCap systems making them compatible with persistent partial occlusions. The "paper" test case represented a generic occlusion created by a random VR device. As conclusion, our method is not only adaptable to MoCap, but it could be also used to generic partial occlusions created by different VR HMD's.

### 5.3.2 MoCap VR method: Assessing Facial Expressions

We divided the validation of our prediction methods in two steps: (i) statistical validation and (ii) visual validation. This section presents the statistical validation of



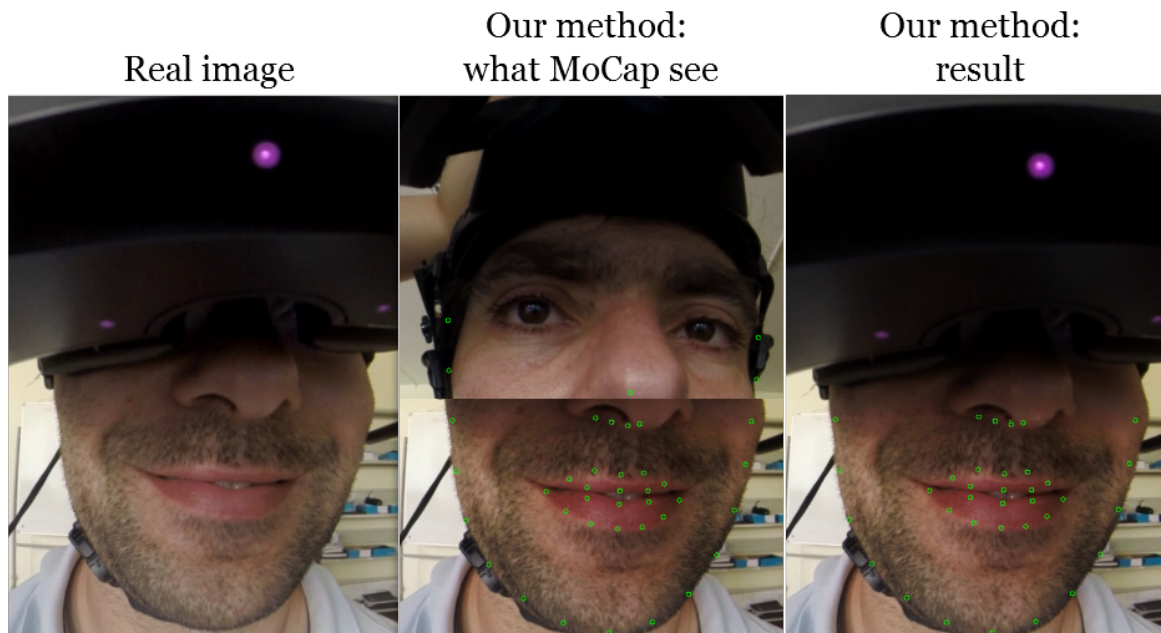


Figure 5.8: MoCap VR method results: Persistent Partial Occlusions method applied to Saragih *et al.* [SLC11a] MoCap. The real image (left), our method result and what MoCap processes (middle) and final result from our method (right).



Figure 5.9: MoCap VR method results: Persistent Partial Occlusions method applied to Cao *et al.* [CHZ14] MoCap.



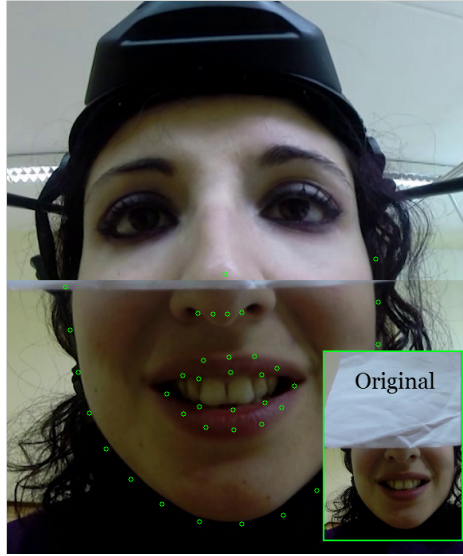


Figure 5.10: MoCap VR method results: Persistent Partial Occlusions method applied to general occlusion created by a paper *et al.* [CHZ14] MoCap.

the methods. To validate our classifiers we adopted a k-Fold Cross Validation (k-Fold CRM) with  $k=10$  [RPL10]. As explained at Chapter 2, k-Fold CRM, after iterating the process of dividing the input data in  $k$  slices for  $k$  times, trains a classifier with  $k-1$  slices. The remaining slices are used as test sets on their respective  $k-1$  trained classifier, allowing us to calculate the accuracy of each one of the  $k-1$  classifiers. The final accuracy value is given by the average of the  $k$  calculated accuracies. Therefore, to each method we analyze k-Fold CRM accuracy to the methods under different scenarios. We highlight that this validation procedure ensures that the test dataset is not the same of the training dataset. Therefore, prediction accuracies are not calculated with test data contained in the training dataset.

Furthermore, we provide a statistical analysis of sensitivity versus specificity and positive versus negative predictive value (i.e. pred. in Tables) [PMP<sup>+</sup>08]. The sensitivity measures the performance of the classifier in correctly predicting the actual class of an item, while specificity measures the same performance but in not predicting the class of an item that is of a different class. Summarizing, sensitivity and specificity measure the true positive and true negative performance, respectively. We added the positive and negative predictive value analysis because these values reflect the probability that a true positive/true negative is correct given knowledge about the prevalence of each class in the data analyzed.

By the end of this section, we used our sample FdMiee database (Chapter 3 - Facial MoCap Tracking) to validate visually our VR methods regarding: occlusions, emotion

and facial expressions prediction. Even though, these data were not acquired with an HMC, the macro expressions captured have high stability regarding head movements. As result, we were able to test our VR method of occlusion and, afterwards, apply and test the facial expressions methods to retrieve information from the upper part of the face.

### 5.3.2.1 MoCap VR Emotion Recognition

Using the k-Fold CRM, we executed a method's validation to two emotion recognition scenarios: (i) six universal emotions of Ekman and Friesen [EF75], plus neutral; (ii) four universal emotions of Jack [JJ13], plus neutral. The six universal emotions [EF75] are the commonly used and accepted by literature studies. However, recent advances in psychology of the emotions show that these emotions are not reproducible throughout different cultures. The non-universality of Ekman's emotions is explored by the survey [JJ13]. This complete study defends that only a subset of the six "universal" emotions is universally recognized, i.e. Joy/Happy, Surprise, Anger and Sad/Sadness. This subset excludes fear and disgust, since these emotions present low recognition cross-culturally being biologically adaptive movements from the emotions surprise and anger, respectively [JJ13].

Therefore, the Table 5.1 shows the k-Fold CRM accuracies to the two scenarios.

Table 5.1: k-Fold CRM Accuracy comparison to scenario (i) and to the scenario (ii). Results in percentage (%).

Emotions	K-Fold Accuracy (%)	95% Confidence Interval
six [EF78]	64.80	(61.72,67.79)
four [JJ13]	69.07	(65.59,72.40)

In the Table 5.1, we observe an increase of the accuracy detection when recognizing four emotions, compared to six emotions classification. This result is not surprising, since we are reducing the number of emotions predicted. In addition, we detect that the bottom features of the face allow a weak recognition of face emotions, resulting in accuracies lower than 70%.

More in detail, we report in the Tables 5.2 and 5.3, a statistical analysis of each emotion recognition obtained with Random Forests classifier to scenario (i) and (ii), respectively.

Both statistical analysis resulted in a p-value lower than  $2.2 \times 10^{-16}$  to a significance

Table 5.2: Statistical Analysis of scenario (i) - Results in percentage (%)

	<b>Anger</b>	<b>Disgust</b>	<b>Fear</b>	<b>Joy</b>	<b>Sadness</b>	<b>Surprise</b>	<b>Neutral</b>
<b>Sensitivity</b>	53.15	39.44	26.09	81.29	12.70	59.40	90.80
<b>Specificity</b>	86.55	97.70	95.84	95.17	99.13	96.35	85.39
<b>Positive pred.</b>	40.21	57.14	39.34	75.90	50.00	71.82	75.51
<b>Negative pred.</b>	91.56	95.40	92.62	96.45	94.31	93.81	94.92

Table 5.3: Statistical Analysis of scenario (ii) - Results in percentage (%)

	<b>Anger</b>	<b>Joy</b>	<b>Sadness</b>	<b>Surprise</b>	<b>Neutral</b>
<b>Sensitivity</b>	75.50	77.85	13.80	68.75	80.09
<b>Specificity</b>	76.16	95.14	99.07	98.39	91.34
<b>Positive pred.</b>	45.06	81.46	66.67	88.51	80.44
<b>Negative pred.</b>	92.31	94.00	89.52	94.59	91.16

level of 5%, which validates our method’s hypothesis: classifying the six/four universal emotions using bottom of face features tracking. Specifically, to scenario (i) at the Table 5.2, we observe an overall low sensitivity to emotions classified (with exceptions to Joy/Happy and Neutral). The opposite is observed to specificity. This indicates that the method does not have high accuracy to detect a certain class, however, does not predict incorrectly. The predictive values weighted using information about the class prevalence in population, show an overall increase of accuracy for true positive and maintain to negative. Therefore, as example to Surprise, despite our classifier only being able to positively identify surprise in 59.40% of the time there is a 71.82% chance that, when it does, such classification is correct. Looking to Table 5.3, compared to previous results of scenario (i) at Table 5.2, we observe an increase of sensitivity, while maintaining an high accuracy of specificity. In general, the same is observed in positive and negative predictive values. This is expected, since decreasing the number of classes of emotions will decrease the degree of confusion that lead to a better split between classes, resulting in a better emotion recognition method. These results confirm the statement of Background section, i.e. bottom face features provide incomplete information about face expression of emotions. Though, our method presents better performance when four universal emotions [JJ13] are classified.

### 5.3.2.2 MoCap VR Facial Expressions Predictor

To analyze and validate the VR facial expressions predictor, we executed the k-Fold cross-validation to the classifier eyebrows ”rising” and to classifier eyebrows

”descending”. Taking into account the variance of nose tracking with the type of HMD used, we propose to study the influence of tracking these features (subset  $S1$ ) and not tracking the nose features (subset  $S2$ ) in the prediction of eyebrows’ movements. Average K-Fold CRM accuracies and respective confidence intervals can be accessed in the Table 5.4.

Table 5.4: k-Fold CRM Accuracy comparison facial expressions assessed (Eyebrows Up or Down) with subset  $S1$  and  $S2$ . Results in percentage (%).

<b>Eyebrows movements</b>	<b>K-Fold Accuracy(%)</b>	<b>95% Confidence Interval</b>
<b>Up <math>S1</math></b>	91.47	(89.76,92.98)
<b>Up <math>S2</math></b>	87.02	(84.97,88.89)
<b>Down <math>S1</math></b>	70.63	(67.99,73.18)
<b>Down <math>S2</math></b>	69.13	(66.40,71.76)

In the Table 5.4, we observe a small decrease of accuracy when the nose features tracking is removed. Although, the confidence intervals show that this decrease is only significant in eyebrows ”up” detection. Our method allows an high performance of eyebrows ”up” estimation (at least, 85%) compared to eyebrows ”down” estimation (at least, 66%). The different results arise from the fact that we are using an emotion database for training, where there is more data describing the ”rising” movement than the opposite (i.e. only anger and sadness emotions usually present this facial expression behavior [EF78]).

Similarly to emotion recognition method, we present the statistical analysis of sensitivity/specificity and positive/negative predictive values to both eyebrows movements using the subsets  $S1$  and  $S2$ .

Table 5.5: Eyebrow Up prediction - Statistical Analysis to subsets  $S1$ . Results in percentage (%).

<b>Eyebrows Up</b>	<b><math>S1</math></b>	<b><math>S2</math></b>
<b>Sensitivity</b>	97.34	96.27
<b>Specificity</b>	71.79	59.18
<b>Positive pred.</b>	92.04	87.65
<b>Negative pred.</b>	92.31	84.06

Both p-values of further analysis are lower than the significance level (i.e. p-value equal to  $2.2 \times 10^{-16} < 0.05$  ). Therefore, both methods are suitable for eyebrows movement estimation using bottom face’s movements. Table 5.4 shows that the method is able to classify the eyebrows ”up” movement accurately, with exception for specificity using the subset  $S2$ . So, the removal of nose features tracking leads, essentially, to a decrease

in accuracy of the classifier in not giving incorrect predictions. However, when we take in to account the prevalence of the class in population, the overall accuracy of prediction to both positive and negative values increase, presenting values above 84.04%.

Table 5.6 contains the statistical analysis to the prediction of eyebrows "descending" movement with (*S1*) and without (*S2*) nose features tracking.

Table 5.6: Eyebrow Down prediction - Statistical Analysis to subsets *S1*. Results in percentage (%).

<b>Eyebrows Down</b>	<b><i>S1</i></b>	<b><i>S2</i></b>
<b>Sensitivity</b>	77.13	73.18
<b>Specificity</b>	62.73	63.97
<b>Positive pred.</b>	71.57	72.09
<b>Negative pred.</b>	69.28	65.23

Observing the Table 5.6, we observe that our method predicts correctly the "descending" movements of the eyebrows, at least, 73.18% of the time and does not predict incorrectly this movements in at least, 63.97% of the time. The lower values are obtained to the subset *S2*, however, the differences between subsets performance are not significant. Similar behavior is beheld taking into account the prevalence of the class in the population. The positive/negative predictive values are not significantly different between sensitivity/specificity. As expected by previous k-Fold CRM results, prediction of the "descending" movement presents lower performance compared to prediction of the opposite movement. Again, this result occurred due to the low prevalence of the "down" class in population. This statement is confirmed by the lower influence shown in positive and negative predictive values when compared to sensitivity and specificity, respectively.

Summarizing, our methods of facial expressions prediction are suitable for the estimation of eyebrows movements using features from the bottom of the face, specially in estimation of the "rising" movement. This conclusion corroborates the hypothesis of this work: our results traduce a connection between bottom and upper face behaviors.

### 5.3.2.3 MoCap VR Assessing Facial Expressions: Visual Results

Applying the VR methods to FdMiee database regarding macroexpressions, we are able to check visually the performance of the methods: occlusions support, emotion recognition and expressions prediction. We chose a non-VR scenario in order to

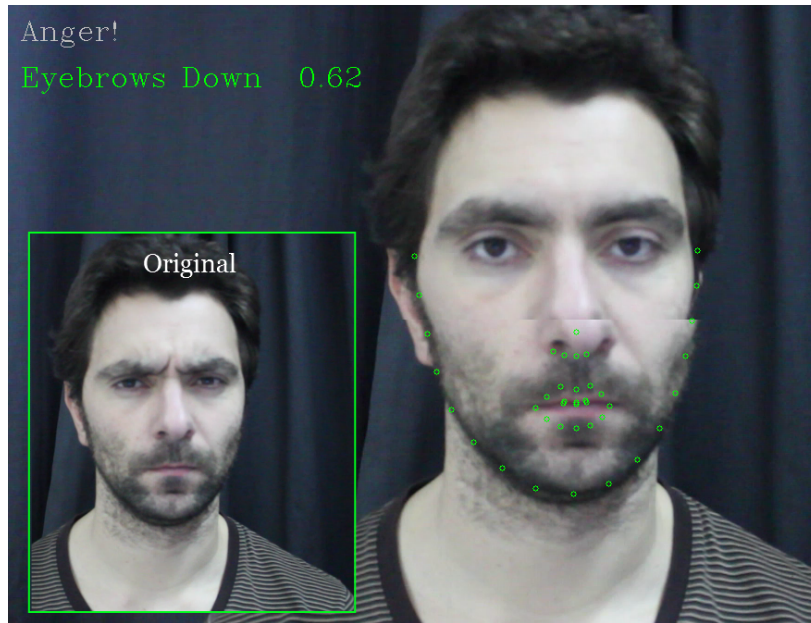


Figure 5.11: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (green). Check that our emotion and prediction match original image eyebrows movements (green box).

verify if the upper face movements and emotions predicted (using only bottom face's movements) match the original facial expressions. Results can be observed in the Figures 5.11, 5.12, 5.13 and 5.14.

Looking throughout the Figures, we verify that our occlusion method is able to "re-create" the face even not using a HMC. Regarding emotion recognition using only the facial features (green dots), in the Figure 5.11, 5.12 and 5.13, we show three examples of correct classification. Figure 5.14 presents an example of a wrong emotion recognition. The classifier returned Anger when the user's emotion label of the video was Sad. This confusion is predicted since the bottom features inherent to Anger and Sad emotions are identical [EF75].

Regarding the facial expressions prediction method, in the Figures 5.11 and 5.14 we observed that the algorithm correctly estimates eyebrows "down", which is confirmed by the original images. The same is detected in the Figure 5.12 for eyebrows "up" predictor. Moreover, in the Figure 5.13, comparing eyebrows of image analyzed and original image, we observe no movement, which traduced in a correct no estimation of movement from both predictors.

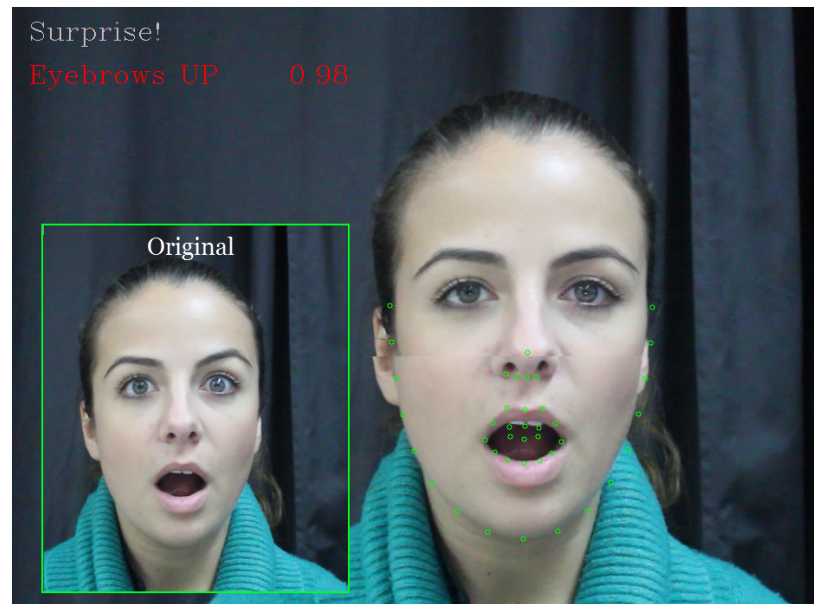


Figure 5.12: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (red). Check that our emotion and prediction match original image eyebrows movements (green box).



Figure 5.13: VR Assessing Facial Expressions: Correct Emotion Recognition result (blue) and no Expression Predictor result, since there is not movement. Check original image in green box.



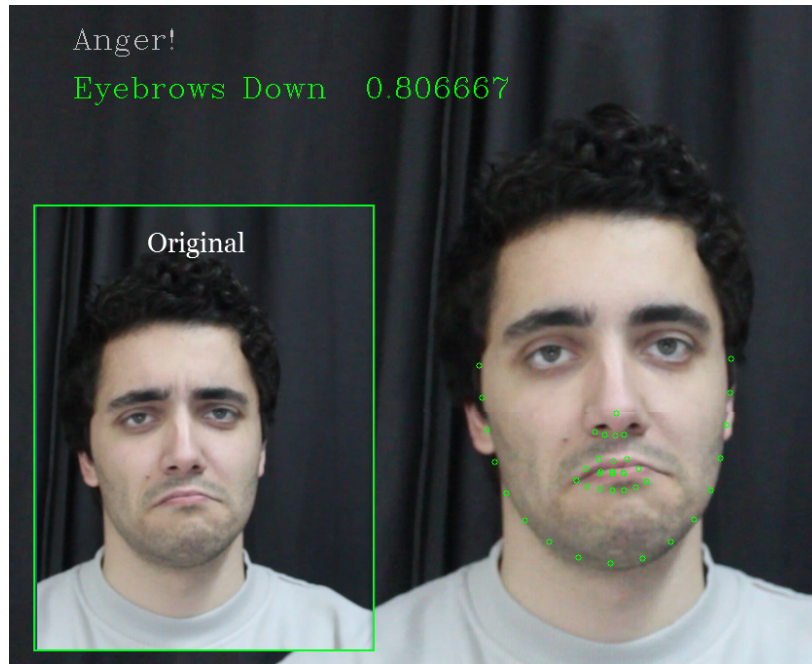


Figure 5.14: MoCap VR Assessing Facial Expressions: Incorrect Emotion Recognition result (blue) and Expression Predictor result. Check original image to see that Expression Predictor is correct (green box).

## 5.4 Conclusions

This chapter presents MoCap VR methods that achieve the three goals: make MoCap systems compatible with persistent partial occlusions, real-time recognition of universal emotions and prediction of upper face movements using bottom face features tracking. The development of these methods lead to improve the three components of sense of embodiment, i.e. enhances the sense of self-location, agency and body ownership within the VR environments [KGS12].

Analyzing the results, we conclude that the three goals proposed where achieved. We deliver a method to make MoCap systems able to track bottom face features under partial occlusions created by different HMD's. Note, we do not deliver a method that is able to overcome generic and unpredicted facial occlusions, since we require the knowledge of which area is occluded. Then, using these facial features, we were able to define methodologies to real-time recognition for four universal emotions (Anger, Joy, Sadness and Surprise) [JJ13], plus Neutral with an accuracy of 69.07% and prediction of facial movements in the occluded regions, i.e. eyebrows "rising" with accuracy of 91.47% and "descending" with an accuracy of 70.63%. The results obtained with the facial expressions prediction method confirmed our method's hypothesis



(see section 5.2 - Methodology). Therefore, besides bottom features of the face being not enough to describe the six emotions of Ekman and Friesen [EF75], our predictor of facial expression decode a connection between bottom face and upper face features. As explained in methodology, the combination of both emotion and expressions tracked/predicted make us able to access a wide range of facial expressions enabling us to represent the diversity of faces [McC93]. This conclusion opens new lines of research to predict more complex movements of the face, even when we are not able to track them using CV algorithms. Furthermore, our methods outputs enable the real-time animation of 3D characters, since we deliver information of facial features combined to emotions that can be used to activate different types of rigs. Ahead of 3D characters animation, our methods are suitable for emotion-based applications, like affective virtual environments, advertising or emotional gaming. In the future, we also open facial animation possibilities to the definition of VR mapping algorithms that use movements and emotions estimated to trigger facial animation compatible with VR environments.



# Chapter 6

## Conclusions and Future Directions

*This PhD thesis has as main motivation the facilitation of non-verbal communication in virtual environments through the automation of 3D characters' facial animation processes. Therefore, we propose, describe and validate novel methods for non-intrusive markerless facial expressions tracking using off-the-shelf hardware and 3D characters animation on-the-fly. In general, we reduce experts' manual intervention and calibration requirements of MoCap facial animation, delivering methods that are usable by anyone and for everyone.*

*In this chapter, we summarize and highlight the main conclusions of each method presented throughout the thesis. By the end, we discuss the future directions and pathways opened by this dissertation. For detailed description of each research topic consult the conclusion of each chapter.*

### 6.1 Conclusions

Believable reproduction of facial movements in 3D characters is still a challenge in CG. To avoid the generation of animation from scratch, several MoCap technologies have been developed [Lew06]. However, high quality results are only possible with skilled artists' fine tuning (e.g. subtle movements inside the lips) or by combination of expensive hardware with tedious user dependent calibrations [ARL<sup>+</sup>09, AFB<sup>+</sup>13, vdPJD<sup>+</sup>14]. These limitations make current MoCap approaches not suitable for non-expert users.

This PhD thesis creates methods to reduce user-dependent calibrations of MoCap

facial animation processes. Adopting non-intrusive and markerless tracking using off-the-shelf hardware, we allow the non-expert users to capture the unique facial traits of a person and map these facial movements into a 3D character, producing animation. We deliver methods in face perception and facial emotions recognition (Chapter 2), facial MoCap tracking and mapping to create on-the-fly animation compatible with virtual environments and VR scenarios (Chapter 4). Furthermore, we deploy modular methods that can be integrated in standalone applications or combined in MoCap facial animation pipelines.

In the MoCap Fundamental Science chapter, we start by investigating the distortions in our own face representation to decode how we perceive our own facial features. As a result, we prove that there is a strong and independent convergent evidence that aspect ratio is the major variation between face representation and real shape. Consequently, humans have poor knowledge about his own face shape with a clear difference between upper and lower face features distribution. Since we study only static facial features, this experiment only allows us to access information about our self-perception of facial morphologies. Thus, we retain that humans are not able to accurately describe their own face shape using static features. Literature show us that we are experts in pointing errors when we observe moving faces [Mor70, Gel08]. Combining this last statement, with perception’s experiment conclusions, we raise the hypothesis: humans hold higher expertise in face perception of facial behaviors than morphologies, i.e. for humans may be easier to recognize and describe faces when they are moving. In the second study of MoCap Fundamental Science, we study the facial behaviors inherent to universal emotions [EF75]. We deliver a real-time geometric method for markerless facial features extraction from faces captured using off-the-shelf hardware. The geometric feature extraction method made possible the creation of a machine learning model that predicts universal emotions, plus neural, with an accuracy of 94%. The emotion recognition experiment was crucial to understand feature definition and extraction methodologies to characterize facial behaviors. Both facial morphologies and behaviors conclusions achieved in fundamental science experiments provided the baseline knowledge for definition and deployment of further chapters’ methods.

To capture the aforementioned unique facial traits (i.e. morphologies and behaviors) of a person, we developed a real-time markerless MoCap tracker (Chapter 3). Our method allows the recognition and tracking of facial features, like cheeks movements, not contemplated by literature algorithms [CHZ14, CWLZ13, LYYB13a, SLC11a]. In addition, we adopt a non-intrusive markerless capture using off-the-shelf hardware and reduce user-dependent calibrations allowing the method’s usage by non-expert users.

The MoCap Tracking method delivered can be used for applications, like behaviors and biometric tracking [LBFCO12] or used as movements tracker for facial animation (Chapter 4). Then, we devise an algorithm that maps the movements tracked by MoCap trackers (like the one delivered in Chapter 3) to 3D characters on-the-fly. Since we adopt a hybrid rigging approach, the mapping method is able to generate animation without previous knowledge of the movements enabling the reproduction of facial traits of the person, like an asymmetric movement of the mouth. The mapping method is independent of the tracker, making it possible to adapt to the user choice MoCap tracking method. Combining the tracking and mapping methods, we come up with a MoCap facial animation methodology that is easy to use and does not requires expensive hardware and complex markers placement.

The increase of availability of consumer-level VR headsets motivated the development of Chapter 5 methods. We deployed methods for MoCap under persistent partial occlusions produced by VR headsets. With our VR methods, we make possible the facial tracking of bottom face region, recognition of universal emotions [EF75] and estimation of eyebrows movements. As a result, we enable the capture of users' face movements and expressions in VR environments, for mood-based applications to access the data and produce customized feedback, like in emotional gaming (Left 4 Dead 2 by Valve) or psychology based experiments [MSSVT15].

To evaluate and validate our MoCap tracking and VR methods (Chapter 3 and 5), we generated the FdMiee protocol to simulate and capture lifelike scenarios, like extreme lightning conditions, and created a sample database. Through the manipulation of FdMiee parameters, the user can simulate a wide range of environment and expressions to test generic CV algorithms.

The contributions of this PhD thesis were validated through the publication of three journal papers and two ACM articles (see Appendix A, B, C, D, E and F). Furthermore, these methods were presented and included in several applications of: EU projects, such as VERE [VER10] and GOLEM [GOL] and national projects, like LIFEisGAME [AMQO13, LIF09]. The technology was presented and reviewed by an international panel of experts from the EU commission. More recently, we also submitted an education based experiment that uses our MoCap tracking and emotion recognition methods to access learners mood during problem-solving situations (Appendix G).

## 6.2 Future Directions

When we look back to what was done, we always feel like we could have done much more and better. The main limitation found in the methods delivered by this PhD thesis is the limited number of user-based testing and evaluation. Bellow is a description of some interesting directions for future research that can benefit from our methods and even extend them:

- **MoCap Fundamental Science:** We raised questions regarding the human self-perception of facial behaviors directly connected to the Uncanny Valley research topic [Mor70, Gel08]. To answer these questions, we encourage the setup of novel experiments in human perception of 3D characters. In the second study, we deliver a real-time emotion recognition method that can be included in several mood-based algorithms with applications in a wide range of fields, like: psychology studies, emotional gaming or advertising.
- **Facial MoCap Tracking:** Using the features tracked by our MoCap method, we look forward to the research of customized model-free trackers. We define these methods as stabilized optical flow trackers where the landmarks can be customized by the user to fit the application. For example, a user with a peculiar wrinkle in the forehead that plays a crucial role in the way he expresses emotions. With tracker customization, we are able to capture the wrinkle movements and transfer to a 3D character improving the believability of the movements. The wide range of features tracked by our method and non-expertise requirement for method's usage, enable the setup of more MoCap fundamental science experiments to decode which are the key features that allow the description of the diversity of human facial behaviors. Besides the triggering of facial animation, the MoCap method modularity encourages us to explore their usage in real-time applications, such as biometric measures [LBFCO12], security with facial traits identification [CI99] and adapt to VR scenarios using the knowledge acquired in the Chapter 5.
- **MoCap Facial Animation:** Using our mapping method, we encourage researchers and developers to execute tests with more trackers and 3D character's rigs and observe the impact in facial animation realism. In addition, we deliver a facial animation system for non-experts, like therapists and teachers, making possible and promoting the definition of new education and psychology research topics [AMQO13, LIF09].

- **MoCap VR Methods:** Consumer-level VR applications were only introduced recently. Therefore, VR research at this level is still an open topic with unlimited potentialities. With our VR methods, we allow the MoCap tracking for the user embodiment in VR applications. For example, we promote the exploration of the following potentialities of our VR methods: to create facial animation in VR scenarios; to develop mood-based VR applications, where user's emotions and facial expressions trigger different feedbacks and responses from the application; to study how the VR MoCap information can increase the user's immersivity [KGS12, SSV14] and setup psychology based experiments for educational or therapeutic purposes.

### 6.3 Take Home Message

The human expertise and dependence in recognizing and using facial expressions to communicate and the diversity of faces across cultures makes the facial MoCap and animation a hot topic in both research and industry. Besides faces' complexity, the way we perceive facial expressions in 3D characters is still unclear. The impact of developing tools that minimize expertise requirements in facial animation makes possible the setup of novel experiments by professionals from other fields or even by the general user. For example, we enable the execution of experiments in psychology and human perception or the collection of big data from users. The data generated will allow to explore how humans' evaluate the face's morphologies and behaviors to improve realism in *MoCap facial animation*. In addition, we encourage scientific community to explore the potentialities of facial MoCap tracking and animation in decoding which are the facial traits that influence our notions of beauty and visual likeness or behind processes, like aging and health status. The discoveries made in these studies may return new features definitions igniting the development of customized facial MoCap systems. Thus, this PhD thesis provides complex algorithms compiled as simple solutions creating the perfect baseline to the next generation of interactive and virtual communications. As a final remark we strongly believe that this PhD thesis opens a new line of research for CG, CV, psychology and VR embodiment for the creation of new applications and new science that before was not possible.





## Appendix A

# Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation

# Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation

Christina T. Fuentes<sup>1</sup>, Catarina Runa<sup>2</sup>, Xenxo Alvarez Blanco<sup>2</sup>, Verónica Orvalho<sup>2</sup>, Patrick Haggard<sup>1\*</sup>

<sup>1</sup> Institute of Cognitive Neuroscience, University College London, London, United Kingdom, <sup>2</sup> Instituto de Telecomunicações, Porto Interactive Center, Universidade do Porto, Porto, Portugal

## Abstract

Despite extensive research on face perception, few studies have investigated individuals' knowledge about the physical features of their own face. In this study, 50 participants indicated the location of key features of their own face, relative to an anchor point corresponding to the tip of the nose, and the results were compared to the true location of the same individual's features from a standardised photograph. Horizontal and vertical errors were analysed separately. An overall bias to underestimate vertical distances revealed a distorted face representation, with reduced face height. Factor analyses were used to identify separable subconfigurations of facial features with correlated localisation errors. Independent representations of upper and lower facial features emerged from the data pattern. The major source of variation across individuals was in representation of face shape, with a spectrum from tall/thin to short/wide representation. Visual identification of one's own face is excellent, and facial features are routinely used for establishing personal identity. However, our results show that spatial knowledge of one's own face is remarkably poor, suggesting that face representation may not contribute strongly to self-awareness.

**Citation:** Fuentes CT, Runa C, Blanco XA, Orvalho V, Haggard P (2013) Does My Face FIT?: A Face Image Task Reveals Structure and Distortions of Facial Feature Representation. PLoS ONE 8(10): e76805. doi:10.1371/journal.pone.0076805

**Editor:** Marcello Costantini, University G. d'Annunzio, Italy

**Received:** June 27, 2013; **Accepted:** September 3, 2013; **Published:** October 9, 2013

**Copyright:** © 2013 Fuentes et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This research was supported by EU FP7 project VERE 257696, work package 1. PH was further supported by an ESRC Professorial Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

\* E-mail: haggard@ucl.ac.uk

## Introduction

Face perception is a central topic in modern psychology. The field has overwhelmingly used visual stimuli and focussed on face recognition, even when considering perception of one's own face [1]. People see their own face only rarely – vanishingly rarely until the recent ready availability of mirrors. Nevertheless, several studies indicate a specific mechanism involved in recognising one's own face (e.g., [2], see also 3 for a review). Much of this literature has on focussed sensitivity to facial symmetry and its relation to effects of mirrors [4,5], and cerebral hemispheric specialisation [6]. Many visual face recognition studies suggest a superior and accurate visual representation of one's own face [3]. However, the persistence of this advantage even when faces are inverted suggests that it relies on local rather than configural processing [7].

In general, the self-face visual recognition literature cannot readily distinguish between self-face processing based on familiarity with a visual image of one's own face suitable for template matching, or based on structural knowledge about what one's face is like (i.e., a face image or a hypothetical stored representation containing information about the positions of facial features relative to one another, akin to the

body structural description [8]. Here we largely remove the visual recognition aspect of self-face processing to focus on the latter, structural representation aspect. Only one study has investigated somatosensory self-face perception [9], and found generally poor performance. Therefore, it remains unclear what people know about their own facial structure, and how this knowledge is stored and represented independent of a specific visual stimulus.

We recently developed tasks for studying the sensed position of body parts (Longo and Haggard, 2012), and stored models of one's own body [10,11]. These representations both showed systematic patterns of distortion, which potentially indicate how spatial information about bodies is represented and stored in the brain. Here we report results on representation of one's own facial features using a method that does involve visual recognition. We show, first, that people make large errors in locating their own facial features, particularly underestimating face height. Second, we show through factor analysis that the representation of facial feature locations follows a characteristic structure. The patterns of localisation errors showed covariance across specific subsets of features, which may be relevant to identifying the organisation of face representation at a supra-featural, or

configural level. The overall structure of face representations implies an important distortion of face shape. Our work provides a novel and systematic approach to a classic question of Gestalt psychology: how are configurations of multiple features represented in the brain as a composite pattern? Our results may also be relevant to the considerable concern regarding one's own facial structure and appearance in some individuals and cultures.

## Methods

### Ethics Statement

All participants gave informed written consent. All experiments were approved by the local ethics committee at University College London.

Participants were seated in front of a computer screen in portrait orientation (Dell model 2007 WFPb, measuring 43.5 cm vertical, 27.5 cm horizontal) which displayed only a small central dot. The position of the dot on the screen was randomised across trials. Participants were instructed to imagine their own face projected frontally, life-size on the screen, with the tip of the nose located at the dot. They used a mouse to indicate the locations corresponding to 11 landmark facial features. The figure reproduced as Figure 1A was shown to participants before the experiment to indicate the exact anatomical landmarks intended. Before each trial, a text label (e.g., "bottom of chin", "centre of left eye") briefly appeared centrally on the screen. Environmental lighting was controlled so that they could not see any reflection of their face on the screen. Each landmark was judged five times in a random order. To quantify errors in perceived position of facial features, responses were later compared to the actual locations of those landmarks, obtained by taking a photograph under standardized conditions and rendering it at life-size on the same screen. The average horizontal (x) and vertical (y) error for attempts to locate each facial landmark were calculated.

Fifty participants (24 female, average age 25 years) took part. The x data from left-sided landmarks (ears, nose and mouth edges, eyes) was reflected in the midline, and averaged with the corresponding right-sided landmark. This imposed an assumption of facial symmetry, but reduced the number of dependent variables and avoided possible confusion regarding the terms *left* and *right* in the context of the task. By analysing the pattern of errors, we aimed to investigate the internal stored representation of one's own face.

Finally, a subset of 10 participants were asked to attend for a second session, in which the screen was rotated to landscape mode.

## Results

The average error vectors are shown superimposed on a schematic face in Figure 1. They reveal large overall biases in locating facial landmarks. The anatomical structure of the face is very different in the horizontal and vertical dimensions. The horizontal dimension is characterised by symmetry and homology, while the vertical dimension lacks both these attributes. Therefore, we expected different patterns of error in

the X and Y dimensions, and accordingly analysed each dimension separately. In the horizontal dimension, mouth and eye width are overestimated, while nose width is underestimated. In the vertical dimension, the hairline is represented as lower, and the chin as higher, than their true locations, suggesting that the face is represented as shorter than its true height. No simple geometric distortion can explain the *overall* pattern of biases: for example, the compression of face height may appear to be a regression of judgement towards the mean defined by the anchor point on the nose tip. However, eye and ear vertical positions appear to be unaffected by this bias, and the bias is absent in the horizontal dimension, suggesting it is not simply a matter of eccentricity. Moreover, Bonferroni-corrected testing showed significant biases for some facial features close to the anchor point, but not for those farther away (table 1).

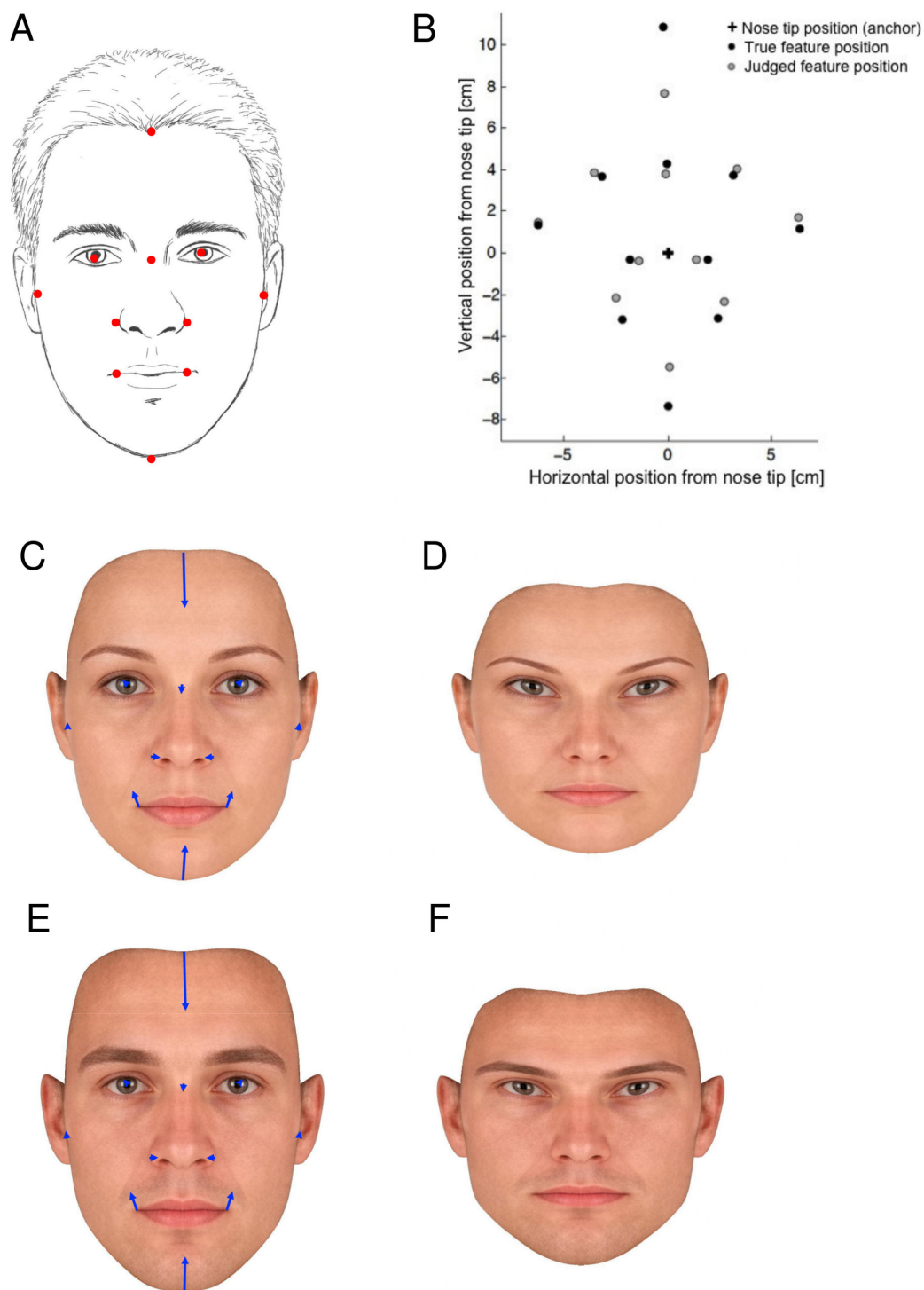
In the ten participants who performed the task with the screen in portrait and landscape mode, we found no effects of screen orientation on judgement error, and no interaction between screen orientation and feature judged, in either X or Y dimensions (all  $F < 1$ , all  $p > 0.60$ ).

To investigate the underlying *structure* of the face representation shown in Figure 1, we applied separate factor analyses to x and y judgement errors (tables S1 and S2). The ratio of measurements-to-cases falls within the guideline range for exploratory factor analysis [12]. Principal components were extracted, and varimax rotated. Factors with eigenvalues over 1 were retained (table 2 and Figure S1).

For horizontal errors, we identified three retainable factors, which we label  $X_1$ ,  $X_2$ ,  $X_3$  for convenience, corresponding to the principal, independent sources of variability in horizontal judgement errors for facial features. The first factor ( $X_1$ ) suggested a tendency to expand facial width outward from the midline. It loaded strongly and roughly equally on all lateralised structures (eye, mouth, ear, nose), but not on midline structures (centre of hairline, bridge of nose, chin). The second factor ( $X_2$ ) suggested lateral distortion of the upper face. It loaded largely on the hairline and nose bridge. The third factor ( $X_3$ ) suggested lateral distortion of the lower face, loading almost exclusively on the chin. For analysis of vertical errors, only two factors were retained. The first ( $Y_1$ ) loaded strongly on upper face structures (eyes), including midline structures (nose bridge, hairline), but with some modest negative loading on the chin. This factor suggested a vertical expansion of the face from its centre. The loadings of the second factor ( $Y_2$ ) on lower face structures (mouth, nose edges, chin) suggest a vertical shift confined to the lower face.

We investigated the relation between the factors underlying face representation and our participants' actual facial features, as measured from photos. Since factor  $X_1$  was interpreted as the width of the face, we correlated scores on this factor with the actual ear-to-ear distance. Since factor  $Y_1$  was interpreted as the vertical height of the face, we correlated it with the actual hairline-to-chin distance. We found no associations between represented and actual facial dimensions ( $r = -0.036$  NS and  $0.016$  NS, respectively).

These factor solutions carry important information about the internal structure of horizontal and vertical face representation.



**Figure 1. Biases in face representation.** A Schematic of feature locations used to instruct participants. B. Actual and mean represented locations. C, Average of 50 female faces reproduced with permission from [www.perceptionlab.com](http://www.perceptionlab.com). Blue arrows indicate mean judgement error for each feature. D. Average female face adjusted according to the mean represented locations of our participants. E, F: as for C, D with average of 50 male faces.

doi: 10.1371/journal.pone.0076805.g001

**Table 1.** Average localisation errors for each feature in cm.

Part	Mean Horizontal Error (cm) (SD)	Mean Vertical Error (cm) (SD)
Hairline	-0.0875 (0.2989)	<b>-3.1533 (1.8734)</b>
Chin	-0.0640 (0.3028)	<b>1.8987 (1.6650)</b>
Ear	0.0396 (1.5981)	0.3534 (1.7092)
Nose Bridge	<b>0.0735 (0.1401)</b>	-0.4734 (1.3835)
Nose	<b>0.4995 (0.6141)</b>	-0.0246 (0.5963)
Mouth	-0.3170 (1.0228)	<b>0.9060 (0.9188)</b>
Eyes	-0.2510 (1.0588)	0.2509 (1.4582)

Values that are significantly different from 0 ( $p < .05$ , after Bonferroni correction for 7 tests) are shown in **bold type**.

doi: 10.1371/journal.pone.0076805.t001

**Table 2.** Factor scores for horizontal X and vertical Y components.

Factor	X1	X <sup>2</sup>	X3	Y1	Y2
Eigenvalue	2.75	1.70	1.05	3.09	1.84
Variance proportion	39%	24%	15%	44%	26%
Hairline	-0.00134	0.91808	-0.08091	0.86393	-0.14580
Chin	-0.02570	-0.01205	0.97501	-0.45231	0.76155
Nose bridge	0.09507	0.89391	0.07555	0.89044	-0.14270
Nose edge	0.66612	-0.24710	-0.11494	0.21907	0.77971
Mouth	0.88904	0.09058	-0.14932	-0.17919	0.92808
Eye	0.90951	0.14324	0.01498	0.92128	-0.02805
Ear	0.78575	0.14074	0.23051	0.32930	0.24252

Only factors with eigenvalues over 1 are shown.

doi: 10.1371/journal.pone.0076805.t002

Factors X1, X<sup>2</sup>, Y1 and Y2 all loaded on more than one facial feature. The loading patterns suggest complexes of two or more individual features that group together, and which covary across the face representations of different individuals. By this means, we could identify separable representations of lateral and midline horizontal facial features, and separable representations of upper and lower face vertical structure. The effects of varying each factor on an average face are shown as vectors in Figure S1, and pictorially in Figure S2.

We also investigated the overall geometry of face representation by seeking an inter-domain association between factors affecting horizontal and vertical errors. We used canonical correlation to identify the principal associations between our horizontal factors (X1, X<sup>2</sup>) and vertical factors (Y1, Y2).

The first canonical variate accounted for 48.5% of the variance between the horizontal and vertical factors and was highly significant (Wilks' Lambda 0.506, approximated by  $F(4,92)=9.34$ ,  $p < .001$ ). The standardised weights showed that the canonical variate related X1 (weighting 0.99) negatively to Y1 (-0.85) and positively, though less strongly, to Y2 (0.53). In contrast, factor X<sup>2</sup> made little contribution to this inter-domain association (weighting 0.12), suggesting that it constituted an independent aspect of facial structure. The combination of weightings in the first canonical variate is readily interpretable

as face aspect ratio, or 2D shape. The lateral shift of eyes, mouth edges, ears and nose captured by factor X1 was associated with a downward shift of the hairline and nose-bridge (captured by Y1), and some upward shift of the mouth, nose edges and chin (captured by Y2). That is, the lateral expansion of the face was strongly associated with a vertical compression of towards the face centre, suggesting that the face aspect ratio is the major structural principle of face representation. The second canonical variate explained only 1.7% of the shared variance between factors, and was far from significant ( $p=0.37$ ). Factor X3 was excluded from the inter-domain analysis, as its loading was largely confined to a single feature. However, re-running the analysis with this factor included had only small effects on weightings of inter-domain association and did not change the pattern of inference. Figure 2A shows the vectors associated with the major loadings ( $>0.4$ ) of each factor, adjusted by the factor's weighting in the canonical variate. Figure 2B shows the face images implied by a positive and negative unit score on the canonical variate.

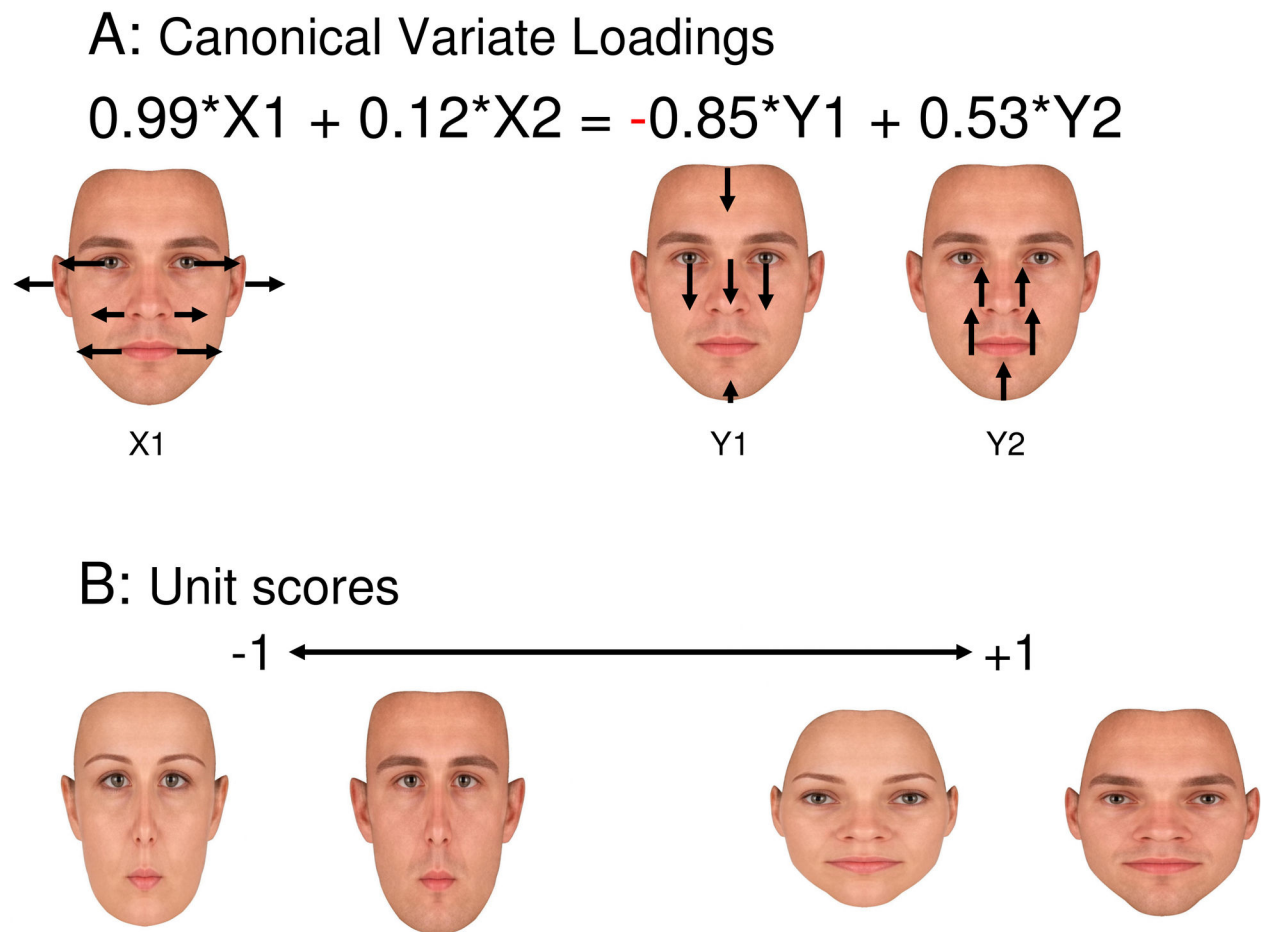
## Discussion

We have developed a new method to investigate stored knowledge about the "face image", or structural arrangement of one's own facial features. Importantly, this method allows the structural description of the face to be investigated independent of visual recognition.

Analyses of errors in locating facial landmarks relative to the tip of the nose suggested an internal representation or model of one's own face, with characteristic structure. We first showed an overall bias to represent face shape as shorter than it really is. This bias was unrelated to the actual height and width of an individual's face. Second, we showed that the most prominent signature of different individuals' overall face representations is the extent to which they express a set of associated factors that code for tall/thin vs short/wide face representation. This recalls similar shape distortions for the position sense of the hand [13], and for the body image [10]. Since shape and size of body parts is not directly signalled by any somatosensory receptor [14], it may be unsurprising that face representation is non-veridical. However, our results show, for the first time, that errors in facial representation are not simply random noise, or regression to the mean, but have a systematic structure.

One striking component of this structure was the aspect ratio defined by facial features. We investigated horizontal and vertical structure of face representation in two independent analyses. We next investigated the association of these dimensions, and found that facial aspect ratio emerged as a prominent feature of the data pattern. Our data therefore provides strong and independent convergent evidence that aspect ratio is a major source of variation in face representation. Not only are people poor at estimating the shape of their own face (Figure 1), but the principal source of variation across individuals is in the biased representation of face shape.

A second clear component of face structure was the separation between upper and lower facial features. For most of the factors we extracted, we found that high loadings on the



**Figure 2. Association between horizontal and vertical distortion factors demonstrates variation in representation of face shape across individuals.** Results of a canonical correlation between the horizontal ( $X1, X2$ ) and vertical ( $Y1, Y2$ ) factors. A. Vectors showing the principal feature loadings ( $>0.4$  or  $<-0.4$ ) of the factors, adjusted by the coefficients indicating important ( $>0.4$  or  $<-0.4$ ) contributions to the canonical variate. The vector lengths are shown at 4x the actual values for visual clarity. Note the negative sign for  $Y1$  coefficient. B. Average female and male faces implied by a low and high score on the canonical variate. Note that the canonical variate separates long and thin from short and wide face representations.

doi: 10.1371/journal.pone.0076805.g002

upper face were accompanied by low loadings on the lower face, or *vice versa*. This dissociation could reflect innervation by different branches of the trigeminal nerve, or it could reflect different functions of the upper face (gaze, attention) and lower face (speech, eating). In any case, our data confirm a fundamental division in face *representation*, as opposed to face perception, between upper and lower face.

Third, we found important misrepresentations of the lateral position of midline structures. Interestingly, these midline shifts occurred independently for the upper face (factor  $X2$ ) and lower face (factor  $X3$ ), providing further strong evidence for independent representation of upper and lower face, but this time from the orthogonal, horizontal dimension of representation. We note that factor  $X3$  requires a more cautious interpretation, given the marginal eigenvalue and loading on a single feature (the chin). The two midline shift

factors could be interpreted as forehead and mandibular asymmetry, respectively. The importance of symmetry in developmental and evolutionary biology is widely accepted [15], and fluctuating asymmetry is also thought to be used as a proxy for biological quality in mate selection [16]. Alternatively, our findings of may reflect brain functions underlying face representation, rather than sensitivity to body morphology. Neuroscientific studies suggest that the two cerebral hemispheres may play different roles in face perception [17]. Variation across individuals in such hemispheric specialization might also explain asymmetric representation of one's own face.

Distortions in face representation have been widely reported in visual perception. For example, one study using adaptation procedures investigated aspect suggested that aspect ratio was a core component of face coding in the human brain [18].



However, those studies did not specifically test for *other* distortions of face coding, apart from shape, and could design only a limited range of stimuli to test dimensions of coding hypothesised a priori. In our approach, by contrast, the key dimensions of face coding emerge from the pattern of participants' responses, rather than by experimenters' choice of stimulus set.

### Configural processing

Models of face perception distinguish between information about individual facial features, and 'holistic' or 'configural' information about spacing between features [19]. Psychophysical studies, for example using the composite face effect, confirm that configural information plays an important role in face perception [20,21], and that this information is processed 'holistically'. However, the structure of the underlying Gestalt or face configuration is not known. Most previous studies have focussed on spatial relations between facial features that are either hypothesised a priori, or motivated by general processing considerations independent of face perception. These include relations between the upper and lower and left and right facial features [17]. In contrast, the Face Image Task (FIT) provides a new, hypothesis-free method for investigating how multiple features are combined in configural representations, at least for representation of one's own face.

In particular, our factor method extracted distinct sets of features whose representations tended to covary, even though we did not impose such a pattern of variation by designing our stimuli, and even though only one feature was ever judged at a time. This grouping of features was not simply defined by proximity (e.g., the edges of the nose grouped with the chin in factor Y2, not with the eyes, despite being closer to the latter than to the former). We suggest that such feature grouping may underlie configural face processing, and could provide a useful data-driven method for identifying what structural information is actually stored in the hypothesised configural representation. Configural processing might reflect precise representation of the spatial relations of features *within* a group, while spatial relations *between* groups of features might be less precisely represented. These findings generate testable predictions for future face-recognition experiments. For example, laterally shifting ears relative to eyes should be readily detectable, due to the common high loadings of these features on factor X1. But vertically shifting ears relative to eyes should be less detectable, since these features are not strongly grouped by any important factor.

### Perceptual and productive self-representation

Our results show that the structural knowledge about one's own facial features is remarkably poor. This contrasts with numerous results in visual self-face recognition showing that self-face processing is remarkably good, and superior to processing of other faces (e.g., [7]). Our results suggest that the internal *representation* of the face is strongly and systematically distorted, but we have no difficulty in recognising much smaller distortions when *viewing* faces (Figure 1). This points to a dissociation between the processes of matching

visual input to a perceptual template, and the processes of accessing structural representations directly for purposes of reproducing them. Artists often improve their face drawing skills by learning geometric rules regarding the spacing of facial features. This may be considered a transfer of training from perceptual representation to productive representation. Interestingly, this process is accompanied by strengthened representation of local featural detail in face perception, at the expense of holistic, configural processing [22,23]. Comparisons of self-face and other-face processing also suggest a dominance of local over configural information for one's own face [24]. Our data suggest that configural information about one's own face is also poorly represented because there are systematic biases in judgements about feature locations. Nevertheless, we found grouping of features in virtue of loading on a single factor. This suggests that some configural structure to face representation is present, albeit of limited accuracy.

In addition, our results offer a dramatic example of the asymmetry between fluent, automatic, stimulus-driven access to object representation, and the limited accessibility of such object representations to the kind of deliberate controlled processing involved in our task. Even our own face appears to be impenetrable to controlled cognition. It is well-known from memory research that recognition is superior to recall. In contrast, the everyday concept of self-awareness implies an opposite pattern. We do not need to recognise our thoughts and mental states as ours. Rather, a stable, persistent core self is held to be directly known, and to provide an origin for mental states, attitudes and actions. This account of the self has recently been questioned [25]. Our approach suggests that bodily self-knowledge is poor, even for elements such as the face, which may be important for personal identity. Therefore, if there is a stable core self underlying self-identity, knowledge about the physical structure of one's own face does not appear to be strongly linked to it.

### Specificity

It is unclear whether the distortions reported here are specific to representing one's own face, or indeed to faces as a category. Identifying suitable objects for a control task is problematic. The quality and quantity of experience we have with other people's faces, and with non-face objects, is entirely different from the experience of our own face. Controlling for modality, familiarity, prototypicality and other relevant factors is therefore difficult. Further, the features of non-face objects cannot match those of faces in number, salience and configuration, almost by definition. Thus, the representation of information about faces cannot easily be compared to representation of other objects. Many perceptual studies suggest a specialised brain system for face processing [26], consistent with specificity. In addition, processing of one's own face may involve a specialised network not used, or used to a lesser extent, for processing of other faces [3]. Comparisons between perception of faces and of non-face objects generally focus on neural *processes*, reflecting the difficulty of comparing the *content* of information represented [27].

For these reasons, it remains unclear if our effects are specific to representations of one's own face. However, the

bias towards short and wide face representation recalls similar biases for hands [11] and body shape [10]. The literature on visual perception and memory for shape do not suggest similar distortions for other objects. For example, people robustly overestimate vertical visual distances compared to horizontal distances [28], whereas we found a striking 27.7% underestimation of face height with relatively unbiased representation of face width (Figure 1). A previous study reported systematic overestimates of one's own head size [29]. However, this conclusion was based on drawing outlines rather than locating features, and more specific analyses identified primarily width overestimation rather than height overestimation [30]. Classic studies of memory for feature locations report several Gestalt-type distortions of spatial representation, but do not mention distortions of aspect ratio [31]. The extensive literature on memory representations for complex figures [32] scarcely mentions distortions of shape – yet it seems unlikely that bias and variability as striking as those we have found for face representation would simply be overlooked. Therefore, we tentatively suggest that the effects reported here may be face-specific, but more research is needed.

### Alternative explanations

Could the factor structure we identified arise artefactually, from some process other than face representation? One possibility is a simple rotational error. Any head tilt in the facial photographs we used to measure judgement accuracy, or in the internal representation of the face that participants used to locate features, would produce systematic errors in judging the features of positions. The misrepresentation of face shape cannot be explained in this way because shape is invariant under rotation. However, some of the other distortions we noted could potentially be due to rotation. Tilt of the head (canting) is particularly likely [33], and is known to influence face recognition [34]. The pattern of errors would depend on the precise centre of rotation. For example, a tilt of the head around the centre of the face would cause equal and opposite X shifts in the hairline and chin. Crucially, our analyses would place these shifts in the same factor, with equal and opposite loadings, because the two shifts are perfectly correlated. In fact, we found that hairline and chin shifts were associated with orthogonal factors. Therefore errors in feature judgements do not appear to be due to face rotation.

A second alternative explanation would involve the spatial distribution of pointing errors around the fixation/anchor point. For example, regression to the mean might cause people to judge all facial features as closer to the nose-tip anchor point than their true location. On this account, errors should vary strictly geometrically with each feature's position in the face, but we found several aspects of face representation that were feature-specific and independent of position in the face or on the screen. For example, we found that errors in localising the bridge of the nose were lower than errors in localising the edges of the mouth (Figure 1), even though both are approximately equidistant from the nose-tip anchor. Our factor analyses confirmed that individual features make distinct contributions to face representation, which are not simply explained by the feature's location within the face. For

example, factor Y2 loaded strongly on the mouth, but much less on the nose edges and chin, even though these features are all close together. Further, simple geometric features of our response method cannot readily explain the strong correlations between factors underlying vertical and horizontal errors. In a previous study of hand representation, patterns of distortion were shown to be invariant when the hand was presented rotated by 90 degrees relative to the body. This suggested the distortion arose from an allocentric representation of the hand, rather than from egocentric or screen-based responding. Such tests can rule out response-specific explanations of bodily distortions for the hand. Such a test is more challenging for face representation, because the face cannot be repositioned within egocentric space in the same way as the hand.

### Limitations

Finally, we acknowledge several limitations of our study. First, the number of participants is small, though it meets standards for exploratory factor analysis based on detailed simulation studies [12]. Second, our data reduction method enforced symmetry of the face around the midline, so is insensitive to possible asymmetries in representation of lateral face structures. Fluctuating asymmetry is an important facial cue to health, genetic quality, and judgements of attractiveness [35]. Future research should examine facial symmetry systematically by testing larger groups, and by directly comparing laterally inverted (mirror) versus confrontational (photograph) representations of the face [36]. Interestingly, we nevertheless identified factors involving midline shifts, confirming that asymmetry is an important aspect of face representation. Third, we have tested location judgement relative to just one central anchor, the tip of the nose. Using another anchor might, in principle, give different results – although tests of body image were largely unaffected by moving the anchor from the head to the feet [10]. Fourth, we tested only the representation of one's own face, so we cannot say whether comparable distortions exist for less familiar faces of others, or for faces as a general semantic category. Fifth and finally, we have used factor analysis to identify the general structure of face representations from individual participants' errors. However, we could not investigate how differences *between* individuals may influence their face representation, due to limited sample size. In particular, an individual's face representation might depend on their actual facial structure, on their gender, or on cultural factors such as a desire to play down unusual or "unattractive" features.

### Supporting Information

**Table S1. Correlation matrix for horizontal errors in feature localisation.**  
(DOCX)

**Table S2. Correlation matrix for vertical errors in feature localisation.**  
(DOCX)



**Figure S1. Results of factor analysis of the face image task reveal principal factors of horizontal and vertical distortion in face representation, rendered on an average female face.** Vector show the principal feature loadings ( $>0.4$  or  $<-0.4$ ) of each factor. The vector lengths are shown at 4x the actual values for visual clarity. The percentage variance and tentative interpretation of each factor are given.

(TIF)

**Figure S2. Pictorial representation of the principal factors of horizontal and vertical distortion.** For each factor, the upper row shows an average male face distorted by a positive score of 1 standard deviation, and the bottom row shows the same face distorted by a negative unit score. Only features with high ( $>0.4$  or  $<-0.4$ ) loadings on the relevant factor were used to render the distortions.

## References

- Uddin LQ, Kaplan JT, Molnar-Szakacs I, Zaidel E, Iacoboni M (2005) Self-face recognition activates a frontoparietal 'mirror' network in the right hemisphere: an event-related fMRI study. *Neuroimage* 25: 926–935. doi:10.1016/j.neuroimage.2004.12.018. PubMed: 15808992.
- Rooney B, Keyes H (2012) Shared or separate mechanisms for self-face and other-face processing? Evidence from adaptation. *Front. Psychology* 3: 66. doi:10.3389/fpsyg.2012.00066.
- Devue C, Brédart S (2011) The neural correlates of visual self-recognition. *Conscious Cogn* 20: 40–51. doi:10.1016/j.concog.2010.09.007. PubMed: 20880722.
- Brédart S (2003) Recognising the usual orientation of one's own face: the role of asymmetrically located details. *Perception* 32: 805–811. doi: 10.1068/p3354. PubMed: 12974566.
- Brady N, Campbell M, Flaherty M (2005) Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. *Brain Cogn* 58: 334–342. doi:10.1016/j.bandc.2005.01.001. PubMed: 15963384.
- Brady N, Campbell M, Flaherty M (2004) My left brain and me: a dissociation in the perception of self and others. *Neuropsychologia* 42: 1156–1161. doi:10.1016/j.neuropsychologia.2004.02.007. PubMed: 15178167.
- Keyes H, Brady N (2010) Self-face recognition is characterized by 'bilateral gain' and by faster, more accurate performance which persists when faces are inverted. *Q J Exp Psychol (Hove)* 63: 840–847. doi: 10.1080/17470211003611264.
- Corradi-Dell'Acqua C, Hesse MD, Rumiati RI, Fink GR (2008) Where is a nose with respect to a foot? The left posterior parietal cortex processes spatial relationships among body parts. *Cereb Cortex* 18: 2879–2890. doi:10.1093/cercor/bhn046. PubMed: 18424775.
- Casey SJ, Newell FN (2005) The role of long-term and short-term familiarity in visual and haptic face recognition. *Exp Brain Res* 166: 583–591. doi:10.1007/s00221-005-2398-3. PubMed: 15983771.
- Fuentes CT, Pazzaglia M, Longo MR, Scivoletto G, Haggard P (2013) Body image distortions following spinal cord injury. *J Neurol Neurosurg Psychiatry* 84: 201–207. doi:10.1136/jnnp-2012-304001. PubMed: 23204474.
- Longo MR, Haggard P (2012) Implicit body representations and the conscious body image. *Acta Psychol (Amst)* 141: 164–168. doi: 10.1016/j.actpsy.2012.07.015. PubMed: 22964057.
- Mundfrom DJ, Shaw DG, Ke TL (2005) Minimum Sample Size Recommendations for Conducting Factor Analyses. *Int J Test* 5: 159–168. doi:10.1207/s15327574ijt0502\_4.
- Longo MR, Haggard P (2010) An implicit body representation underlying human position sense. *Proc Natl Acad Sci U S A* 107: 11727–11732. doi:10.1073/pnas.1003483107. PubMed: 20547858.
- Gandevia SC, Phegan CM (1999) Perceptual distortions of the human body image produced by local anaesthesia, pain and cutaneous stimulation. *J Physiol Lond* 514 ( 2): 609–616. doi:10.1111/j.1469-7793.1999.609ae.x. PubMed: 9852339.
- Palmer AR, Strobeck C (1986) Fluctuating Asymmetry: Measurement, Analysis, Patterns. *Annu Rev Ecol Syst* 17: 391–421. doi:10.1146/annurev.es.17.110186.002135.
- Little AC, Jones BC, Burt DM, Perrett DI (2007) Preferences for symmetry in faces change across the menstrual cycle. *Biol Psychol* 76: 209–216. doi:10.1016/j.biopsycho.2007.08.003. PubMed: 17919806.
- Ramon M, Rossion B (2012) Hemisphere-dependent holistic processing of familiar faces. *Brain Cogn* 78: 7–13. doi:10.1016/j.bandc.2011.10.009. PubMed: 22099150.
- Watson TL, Clifford CWG (2003) Pulling faces: an investigation of the face-distortion aftereffect. *Perception* 32: 1109–1116. doi:10.1068/p5082. PubMed: 14651323.
- Piepers DW, Robbins RA (2012) A review and clarification of the terms 'holistic,' 'configural,' and 'relational' in the face perception literature. *Front. Psychology* 3: 559. doi:10.3389/fpsyg.2012.00559.
- Tanaka JW, Gordon I (2011) Features, Configuration, and Holistic Face Processing. In: G RhodesA CalderM JohnsonJV Haxby. *Oxford Handbook of Face Perception*. Oxford University Press.
- Young AW, Hellawell D, Hay DC (1987) Configurational information in face perception. *Perception* 16: 747–759. doi:10.1068/p160747. PubMed: 3454432.
- Chamberlain R, McManus IC, Riley H, Rankin Q, Brunswick N (2012) Local processing enhancements associated with superior observational drawing are due to enhanced perceptual functioning, not weak central coherence. *Q J Exp Psychol (Hove)*. doi: 10.1080/17470218.2012.750678.
- Zhou G, Cheng Z, Zhang X, Wong ACN (2012) Smaller holistic processing of faces associated with face drawing experience. *Psychon Bull Rev* 19: 157–162. doi:10.3758/s13423-011-0174-x. PubMed: 22215464.
- Greenberg SN, Goshen-Gottstein Y (2009) Not all faces are processed equally: evidence for featural rather than holistic processing of one's own face in a face-imaging task. *J Exp Psychol Learn Mem Cogn* 35: 499–508. doi:10.1037/a0014640. PubMed: 19271862.
- Metzinger T (2004) *Being No One: The Self-model Theory of Subjectivity*. Cambridge, MA: The MIT Press.
- Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. *Philos Trans R Soc Lond, B, Biol Sci* 361: 2109–2128. doi:10.1098/rstb.2006.1934. PubMed: 17118927.
- Gauthier I, Skudlarski P, Gore JC, Anderson AW (2000) Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci* 3: 191–197. doi:10.1038/72140. PubMed: 10649576.
- Avery GC, Day RH (1969) Basis of the horizontal-vertical illusion. *J Exp Psychol Hum Learn* 81: 376–380. PubMed: 5811814.
- Bianchi I, Savardi U, Bertamini M (2008) Estimation and representation of head size (people overestimate the size of their head - evidence starting from the 15th century). *Br J Psychol* 99: 513–531. doi: 10.1348/000712608X304469. PubMed: 18471345.
- Savardi U, Bianchi I (2006) Quanto grande è la mia testa? Contributi dalla fenomenologia sperimentale della percezione. *DIPAV - Quaderni* 15: 59–78.
- Tversky B (1981) Distortions in memory for maps. *Cogn Psychol* 13: 407–433. doi:10.1016/0010-0285(81)90016-5.

(TIF)

## Acknowledgements

We are grateful to Dave Perrett and Amanda Hahn of the Perception Lab, University of St Andrews, [www.perceptionlab.com](http://www.perceptionlab.com), for permission to use their average face images.

## Author Contributions

Conceived and designed the experiments: CTF VO PH. Performed the experiments: CTF. Contributed reagents/materials/analysis tools: CTF CR XB VO PH. Wrote the manuscript: CTF PH.

32. Shin MS, Park SY, Park SR, Seol SH, Kwon JS (2006) Clinical and empirical applications of the Rey-Osterrieth Complex Figure Test. *Nat Protoc* 1: 892–899. doi:10.1038/nprot.2006.115. PubMed: 17406322.
33. Costa M, Menzani M, Bitti PER (2001) Head Canting in Paintings: An Historical Study. *J Nonverbal Behav* 25: 63–73. doi:10.1023/A:1006737224617.
34. Collishaw SM, Hole GJ, Schwaninger A (2005) Configural processing and perceptions of head tilt. *Perception* 34: 163–168. doi:10.1068/p5216. PubMed: 15832567.
35. Rhodes G (2006) The evolutionary psychology of facial beauty. *Annu Rev Psychol* 57: 199–226. doi:10.1146/annurev.psych.57.102904.190208. PubMed: 16318594.
36. Thomas R, Press C, Haggard P (2006) Shared representations in body perception. *Acta Psychol* 121: 317–330. doi:10.1016/j.actpsy.2005.08.002. PubMed: 16194527.

Table S1

Hairline	1.0000						
Chin	-0.0440	1.0000					
Ear	-0.0561	-0.0906	1.0000				
Nose bridge	0.6839	0.0502	-0.1945	1.0000			
Nose edge	0.1583	0.0457	0.2980	0.0366	1.0000		
Mouth	-0.1171	0.1189	0.5617	-0.1107	0.5069	1.0000	
Eye	-0.1238	0.0195	0.7027	-0.1635	0.4246	0.8085	1.0000
	Hairline	Chin	Ear	Nose bridge	Nose edge	Mouth	Eye

Table S1. Correlation matrix for horizontal errors in feature localisation

Table S2

Hairline	1.0000						
Chin	-0.4810	1.0000					
Ear	0.1745	0.0123	1.0000				
Nose bridge	0.6928	-0.4304	0.2000	1.0000			
Nose edge	0.0811	0.3254	0.0165	-0.0109	1.0000		
Mouth	-0.2733	0.7768	0.1487	-0.2501	0.5842	1.0000	
Eye	0.7149	-0.3836	0.1689	0.8265	0.1694	-0.2021	1.0000
	Hairline	Chin	Ear	Nose bridge	Nose edge	Mouth	Eye

Table S2. Correlation matrix for vertical errors in feature localisation.

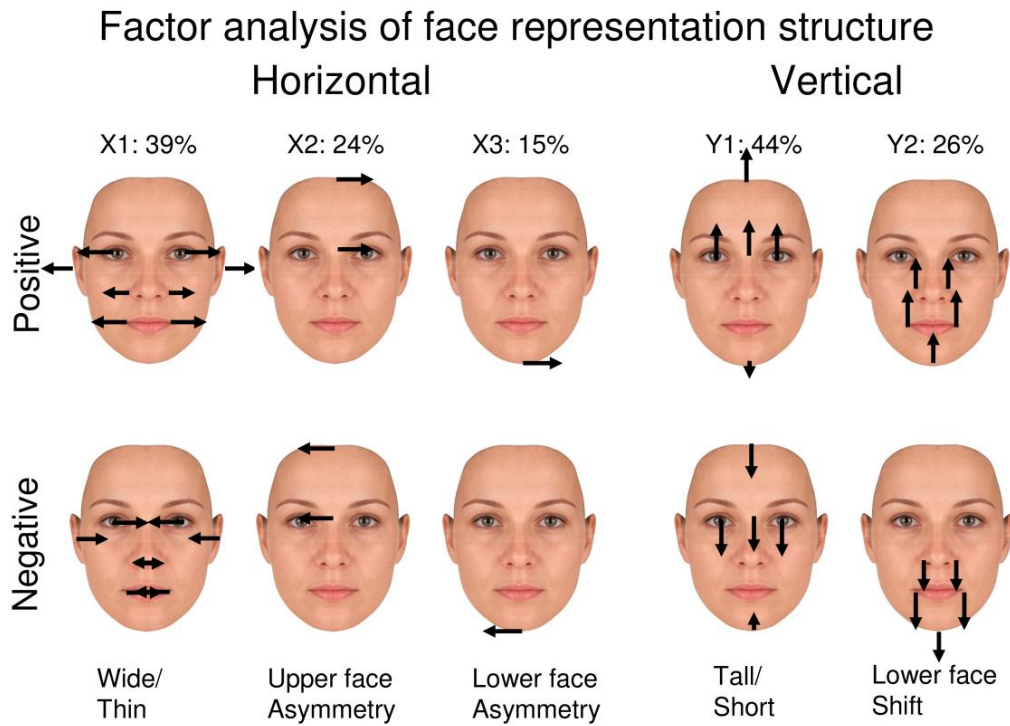


Figure S1. Results of factor analysis of the face image task reveal principal factors of horizontal and vertical distortion in face representation, rendered on an average female face. Vector show the principal feature loadings ( $>0.4$  or  $<-0.4$ ) of each factor. The vector lengths are shown at 4x the actual values for visual clarity. The percentage variance and tentative interpretation of each factor are given

10.1371/journal.pone.0076805.s003

## Factor analysis of face representation structure

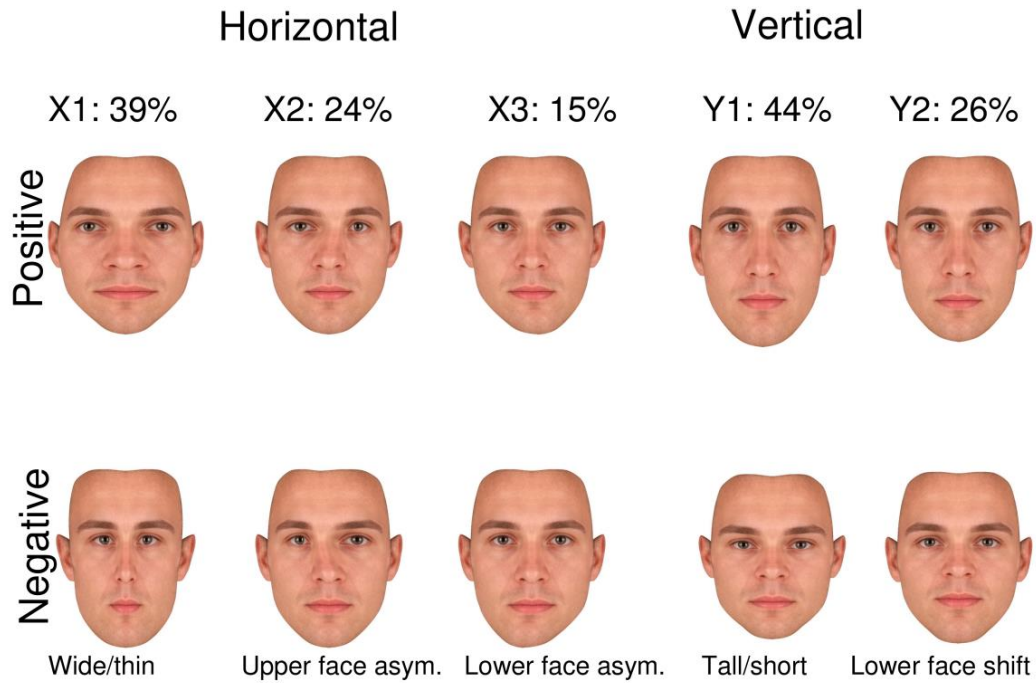


Figure S2. Pictorial representation of the principal factors of horizontal and vertical distortion. For each factor, the upper row shows an average male face distorted by a positive score of 1 standard deviation, and the bottom row shows the same face distorted by a negative unit score. Only features with high ( $>0.4$  or  $<-0.4$ ) loadings on the relevant factor were used to render the distortions

10.1371/journal.pone.0076805.s004



## Appendix B

### Real-Time Emotion Recognition: a Novel Method for Geometrical Facial Features Extraction

# Real-Time Emotion Recognition: a Novel Method for Geometrical Facial Features Extraction

Claudio Loconsole<sup>1</sup>, Catarina Runa Miranda<sup>2</sup>, Gustavo Augusto<sup>2</sup>, Antonio Frisoli<sup>1</sup>, and Verónica Orvalho<sup>2</sup>

<sup>1</sup>*PERCRO Laboratory, Scuola Superiore Sant'Anna*

<sup>2</sup>*Instituto Telecomunicações, Faculdade de Ciências, Universidade do Porto*

**Keywords:** Human-Computer interaction; Emotion Recognition; Computer Vision

**Abstract:** Facial emotions provide an essential source of information commonly used in human communication. For humans, their recognition is automatic and is done exploiting the real-time variations of facial features. However, the replication of this natural process using computer vision systems is still a challenge, since automation and real-time system requirements are compromised in order to achieve an accurate emotion detection. In this work, we propose and validate a novel methodology for facial features extraction to automatically recognize facial emotions, achieving an accurate degree of detection. This methodology uses a real-time face tracker output to define and extract two new types of features: *eccentricity* and *linear* features. Then, the features are used to train a machine learning classifier. As result, we obtain a processing pipeline that allows classification of the six basic Ekman's emotions (plus *Contemptuous* and *Neutral*) in real-time, not requiring any manual intervention or prior information of facial traits.

## 1 INTRODUCTION

Facial expressions play a crucial role in communication and interaction between humans. In the absence of other information such as speech interaction, facial expressions can transmit emotions, opinions and clues regarding cognitive states (Ko and Sim, 2010). A fully automatic real-time face features extraction for emotion recognition allows to enhance the communication realism between humans and machines. There are several research fields interested in developing automatic systems to recognize facial emotions. They mainly are represented by:

- Cognitive Human-Robot Interaction: the evolution of robots and computer animated agents bring a social problem of communication between these systems and humans (Hong et al., 2007);
- Human-Computer Interaction: facial expressions analysis is widely used for telecommunications, behavioural science, videogames and other systems that require facial emotion decoding for communication (Fernandes et al., 2011).

Several face recognition systems have been developed for real time facial features detection as well as (e.g. (Bartlett et al., 2003)). Psychological studies have been conducted to decode this information only using facial expressions, such as the Facial Action Cod-

ing System (FACS) developed by Ekman (Ekman and Friesen, 1978).

As stated on the recent survey (Jamshidnezhad and Nordin, 2012), among existing facial expression recognition systems, the common three-step pipeline for facial expressions classification (Bettadapura, 2009) is composed by:

1. the *Facial recognition* phase;
2. the *Features extraction* phase;
3. the *Machine learning classifier* phase (preliminary model training and on-line prediction of facial emotions).

As claimed in the same survey, the second pipeline phase (features extraction) strongly influences the accuracy and computational cost of the overall system. It follows that the choice of the type of the features to be extracted and the corresponding methods to be used for the extraction is fundamental for the overall performances.

The commonly used methods for feature extraction can be divided into *geometrical* methods (i.e. features are extracted from shape or salient point locations such as the mouth or the eyes (Kapoor et al., 2003)) and *appearance-based* methods (i.e. skin features like frowns or wrinkles, *Gabor Wavelets* (Fischer, 2004)).



Geometric features are selected from landmarks positions of essential parts of the face (i.e. eyes, eyebrows and mouth) obtained by a face features recognition technique. These extraction methods are characterized by their simplicity and low computational cost, but their accuracy is extremely dependent on the face recognition performances. Examples of emotion classification methodologies that use geometric features extraction are (Cheon and Kim, 2009; Niese et al., 2012; Gang et al., 2009; Hammal et al., 2007; Seyedarabi et al., 2004; Kotsia and Pitas, 2007). However, high accuracies on emotion detection usually require a calibration with a neutral face ((Kotsia and Pitas, 2007; Gang et al., 2009; Niese et al., 2012; Cheon and Kim, 2009; Hammal et al., 2007)), an increase of the computational cost ((Gang et al., 2009; Seyedarabi et al., 2004)), a decrease of the number of emotions detected ((Niese et al., 2012; Hammal et al., 2007)) or a manual grid nodes positioning (Kotsia and Pitas, 2007). On the other hand, appearance-based features work directly on image and not on single extracted points (e.g. *Gabor Wavelets* (Kotsia et al., 2008) and *Local Binary Patterns* (Shan et al., 2009) (Chatterjee and Shi, 2010)). They usually analyze the skin texture, extracting relevant features for emotion detection. Involving a higher amount of data, the appearance feature method becomes more complex than the geometric approach, compromising also the real-time feature required by the process (appearance-based features show high variability in performance time from 9.6 to 11.99 seconds (Zhang et al., 2012)). Hybrid approaches, that combine geometric and appearance extraction can be found (i.e. (Youssif and Asker, 2011)) with higher accuracies, but they are still characterized by a high computational cost. The aim of this research work is to propose a feature extraction method that provides performances comparable with appearance-based methods without compromising the real-time and automation requirements of the system. Nevertheless, we intent to solve the following main four facial emotion recognition issues (Bettadapura, 2009):

1. real-time requirement: communication between humans is a real time process with a time scale order of about 40 milliseconds (Bartlett et al., 2003);
2. capability of recognition of multiple standard emotions on people with different anthropometric facial traits;
3. capability of recognition of the facial emotions without neutral face comparison calibration;
4. automatic self-calibration capability without manual intervention.

(equivalent optimizations of these four issues can also be extracted from Jamshidnezhad et al.'s survey (Jamshidnezhad and Nordin, 2012)). *Real-time issue* is solved using a low complexity features extraction

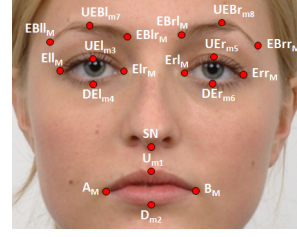


Figure 1: The subset composed by 19 points of the 66 *FaceTracker* facial landmarks used to extract our proposed geometric facial features.

method without compromising the accuracy of emotion detection. In order to show the capacity of the *second issue*, we test our system on a multi-cultural database, the Radboud face database (Langner et al., ), featured with multiple emotions traits (Bettadapura, 2009). Additionally, we investigate all six universal facial expressions (Ekman and Friesen, 1978) (*Joy, Sorrow, Surprise, Fear, Disgust and Anger*) plus *Neutral* and *Contemptuous*. Regarding the *third issue*, though with slightly lower performance relative to neutral face comparison calibration, our method allows the recognition of eight different emotions without requiring any calibration process. To avoid any *manual intervention* in the localization of the seed landmarks required by our proposed geometrical features, we use as reference example in this work, a marker-less facial landmark recognition and localization software based on the *Saragih's FaceTracker* (Saragih et al., 2011b). However, face recognition can be done with the use of different marker-based and marker-less systems which allow the localization of the basic landmarks defined in our system for emotion classification. Therefore, as main contribution, we defined facial features inherent to emotions and proposed a method for their extraction in real time, for further emotion recognition.

## 2 GEOMETRIC FACIAL FEATURES EXTRACTION METHOD

In this work, we propose a set of facial features suitable for marker-based and marker-less systems. In fact, we present an approach to extract facial features that are truly connected to facial expression. We start from a subset composed by 19 elements (see Fig. 1 and Table 1) of the 54 anthropometric facial landmarks set defined in (Luximon et al., 2011) that are usually localized using facial recognition methods.

The testing benchmark used for our extraction method is an existing marker-less system for landmark identification and localization by Saragih et al. (Saragih et al., 2011a). Their approach reduces detec-

Table 1: The subset of anthropometric facial landmarks used to calculate our proposed geometric facial features.

No.	Landmark	Label	Region
1	Right Cheilion	$A_M$	Mouth
2	Left Cheilion	$B_M$	Mouth
3	Labiale Superius	$U_{m1}$	Mouth
4	Labiale Inferius	$D_{m2}$	Mouth
5	Left Exocanthion	$El_{lM}$	Left Eye
6	Right Exocanthion	$El_{rM}$	Left Eye
7	Palpebrale Superius	$UE_{lM3}$	Left Eye
8	Palpebrale Inferius	$DE_{lM4}$	Left Eye
9	Left Exocanthion	$Er_{lM}$	Right Eye
10	Right Exocanthion	$Err_{rM}$	Right Eye
11	Palpebrale Superius	$UE_{rM5}$	Right Eye
12	Palpebrale Inferius	$DE_{rM6}$	Right Eye
13	Zygofrontale	$EB_{lM}$	Left Eyebrow
14	Inner Eyebrow	$EB_{lrM}$	Left Eyebrow
15	Superciliare	$UE_{BlM7}$	Left Eyebrow
16	Inner Eyebrow	$EB_{rlM}$	Right Eyebrow
17	Zygofrontale	$EB_{rrM}$	Right Eyebrow
18	Superciliare	$UE_{BrM8}$	Right Eyebrow
19	Subnasale	$SN$	Nose

tion ambiguities, presents low online computational complexity and high detection efficiency outperforming the other popular deformable real-time models to track and model non-rigid objects (Active Appearance Models (AAM) (Asthana et al., 2009), Active Shape Models (ASM) (Cootes and C.J.Taylor, 1992), 3D morphable models (Vetter, ) and Constrained Local Models (CLMs) (Cristinacce and Cootes, )).

Saragih et al. (Saragih et al., 2011a) system identifies and localizes 66 2D landmarks on the face. Through the repetitive observation of facial behaviours during emotion expressions, we empirically choose a subset of 19 facial landmarks that better capture these facial changes among the 66 *FaceTracker* ones.

Using the landmark positions in the image space, we define two classes of features: *eccentricity* and *linear* features. These features are normalized to the range [0,1] to let the feature not affected by people anthropometric traits dependencies. So, we extract geometric relations among landmark positions during emotional expression for people with different ethnicities and ages.

## 2.1 Eccentricity features

The *eccentricity* features are determined by calculating the eccentricity of ellipses constructed using specific facial landmarks. Geometrically, the *eccentricity* measures how the ellipse deviates from being circular. For ellipses the *eccentricity* is higher than zero and lower than one, being zero if it is a circle. As example, drawing an ellipse using the landmarks of the mouth, it is possible to see that while smiling the eccentricity is higher than zero, but when expressing

surprise it is closer to a circle and almost zero. A similar phenomenon can be observed also in the eyebrow and eye areas. Therefore, we use the eccentricity to extract new features information and classify facial emotions. More in detail, the selected landmarks for this kind of features are 18 over 19 (see Table 1 and Fig. 1), whereas the total defined eccentricity features are eight: two in the mouth region, four in the eye region and two in the eyebrows region (more details can be found in Table 2). Now, we describe the *eccentricity* extraction algorithm applied to the mouth region. The same algorithm can be simply applied to the other face areas (eyebrows and eyes) following the same guidelines.

With reference to Figure 2.a, let  $A_M$  and  $B_M$  be the end points of the major axis corresponding to the side ends of the mouth, while  $U_{m1}$  the upper end points of the minor axis (the distance between the major axis and  $U_{m1}$  corresponds to the semi-minor axis). Of course, the symmetry of  $U_{m1}$  with respect to  $A_M$  and  $B_M$  is not assured. For this reason, in the following, we will refer to each ellipse as the *best fitting ellipse* among the three points having the semi-minor axis equal to the distance between  $U_{m1}$  and the line  $A_MB_M$ .

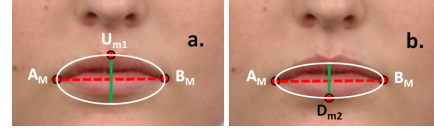


Figure 2: The definition of the first (a.), “upper” and the second (b.), “lower” ellipses of the mouth region using respectively the triple  $(A_M, B_M, U_{m1})$  and  $(A_M, B_M, D_{m2})$ .

We construct the first ellipse  $E_1$ , named “upper” ellipse, defined by the triple  $(A_M, B_M, U_{m1})$  and calculate its eccentricity  $e_1$ . The eccentricity of an ellipse is defined as the ratio of the distance between the two foci, to the length of the major axis or equivalently:

$$e = \frac{\sqrt{a^2 - b^2}}{a} \quad (1)$$

where  $a = \frac{B_{Mx} - A_{Mx}}{2}$  and  $b = A_{My} - U_{m1y}$  are respectively one-half of the ellipse  $E$ ’s major and minor axes, whereas  $x$  and  $y$  indicate the horizontal and the vertical components of the point in the image space. As mentioned above, for an ellipse, the eccentricity is in the range [0,1]. When the eccentricity is 0, the foci coincide with the center point and the figure is a circle. As the eccentricity tends toward 1, the ellipse gets a more elongated shape. It tends towards a line segment if the two foci remain a finite distance apart and a parabola if one focus is kept fixed as the other is allowed to move arbitrarily far away.

We repeat the same procedure for the ellipse  $E_2$ , named “lower” ellipse, using the lower end of the mouth (see Fig. 2.b). The other six ellipses are,

then, constructed following the same extraction algorithm using the features summarized in Table 2 (for the landmark labels refer to Table 1 and Fig. 1). It is clear that for both eyebrows, it is not possible to calculate the lower ellipses due to their morphology. The final results of the ellipse construction can be seen in Figure 3.a, whereas in Figure 3.b it is possible to see how the eccentricities of the facial ellipses changes according to the person’s facial emotion.

Table 2: The eight ellipses used to extract the eccentricity features (for the landmark labels please refer to Fig. 1).

Ellipse	Point Triple	Region
$E_1$	$(A_M, B_M, U_{m1})$	Upper mouth
$E_2$	$(A_M, B_M, D_{m2})$	Lower mouth
$E_3$	$(El_{lM}, El_{rM}, UEl_{m3})$	Upper left eye
$E_4$	$(El_{lM}, El_{rM}, DEl_{m4})$	Lower left eye
$E_5$	$(Er_{lM}, Err_{rM}, UEr_{m5})$	Upper right eye
$E_6$	$(Er_{lM}, Err_{rM}, DEr_{m6})$	Lower right eye
$E_7$	$(EB_{lM}, EB_{lrM}, UEB_{l_{m7}})$	Left eyebrow
$E_8$	$(EB_{rM}, EB_{rrM}, UEB_{r_{m8}})$	Right eyebrow

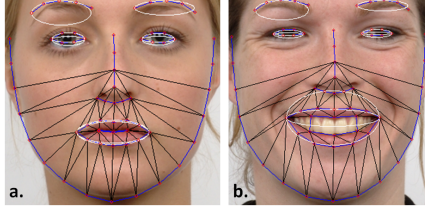


Figure 3: The final results of the eight ellipse construction (a). Eccentricities of the facial ellipses changes according to the person’s facial emotion (b).

## 2.2 Linear features

The *linear* features are determined by calculating linear distances between couples of landmarks normalized with respect to a physiologically greater facial inter-landmark distance. These distances intend to quantitatively evaluate the relative movements between facial landmarks while expressing emotions. The selected distances are those corresponding to the movements between eyes and eyebrows  $L_1$ , mouth and nose  $L_2$  and upper and lower mouth points  $L_3$ . More in detail, with reference to Table 1 and Figure 1, indicating with  $\_y$  only the vertical component of each point in the image space and selecting as  $DEN = \overline{UEl_{m3y}SN_y}$  the normalizing distance, we calculate a total of three linear features as:

1.  $L_1 = \overline{UEB_{l_{m7y}}UEl_{m3y}}/DEN$ ;
2.  $L_2 = \overline{U_{m1y}SN_y}/DEN$ ;
3.  $L_3 = \overline{D_{m2y}SN_y}/DEN$ ;

## 3 EXPERIMENTAL PART

In this Section, we describe the conducted tests to evaluate the emotion recognition performances of the proposed facial geometrical features. More in detail, the *classifier validation* (Section 3.3), is related to investigate three classification methods and select the one that provides the best performances on emotion recognition using both for training and validation a particular subset of proposed features. The *feature evaluation* (Section 3.4), instead, is related to fully evaluate our proposed features using the classification method selected at the end of the first experiment. In Section 3.1, we report the organization of the defined features used in both tests, whereas, in Section 3.2, we illustrate the facial emotion database (the Radboud facial database) used to extract the defined features.

### 3.1 Extracted features

In order to fully evaluate and compare our defined features, we consider five types of feature subsets:

1. only linear features (subset  $S1$ : 3 elements);
2. only eccentricity features (subset  $S2$ : 8 elements);
3. both eccentricity and linear features (subset  $S3$ : 11 elements);
4. differential eccentricity and linear features with respect to those calculated for neutral emotion face (subset  $S4$ : 11 elements);
5. all features corresponding to the union of  $S3$  and  $S4$  (subset  $S5$ : 22 elements).

(where the differential features are calculated as:

$$df_{i,x} = f_{i,x} - f_{i,neutral}$$

with  $i$  representing a subject of the database and  $x$  an emotion), resulting in a total number of calculated features for the entire database equal to  $(1.385 \text{ pictures} \times 22 \text{ } S5 \text{ numerosity}) 30.470$ . The five subsets can be grouped into two main classes:

1. the *intra-person-independent* or *non-differential* subsets  $S1$ ,  $S2$  and  $S3$  that do not require any kind of calibration with other facial emotion states of the same person;
2. the *intra-person-dependent* or *differential* subsets  $S4$  and  $S5$  that require a calibration phase using the neutral expression of the same person.

### 3.2 Database Description

In order to demonstrate the capacity of recognition of multiple standard emotions on people with different anthropometric facial traits, we test our system on a multi-cultural database featured with multiple emotions elements. The selected testing platform is the

Radboud facial database (Langner et al., ). It is composed by 67 real person's face models performing the six universal facial expressions (Ekman and Friesen, 1978) (*Joy, Sorrow, Surprise, Fear, Disgusted and Angry*) plus *Neutral* and *Contemptuous*. Even if the considered images are all frontal, for each couple person-expression, there are three pictures corresponding to slightly different angles of gaze directions, without changing head orientation. This leads to a total of 1608 ( $67 \times 8 \times 3$ ) picture samples.

The pictures are coloured and contain both gender Caucasian and Moroccan adults and Caucasian kids. More specifically, in the database there are 39 Caucasian adults (20 males and 19 females); 10 Caucasian children (4 males and 6 females); 18 Moroccan male adults. Therefore, using this database we provide emotion expressions information relative to a population database that includes gender, ethnic and age variations combined with diverse facial positioning. This will allow us to create a model that will predict emotion expression even with this diverse changes.

To decouple the performances of our method's validation (in the scope) and those of the *FaceTracker* software (out of the paper scope), we adopt a pre-processing step. During this pre-processing we removed 223 elaborated picture samples in which the landmarks were not properly recognized by *FaceTracker* software, leading to a total number of tested pictures equal to 1385. With this outlier removal, we guarantee a correct training of the machine learning classifier, since we capture correctly the facial behaviors inherent to considered emotions.

### 3.3 Classifier validation

The *classifier validation* test is subdivided into two parts:

1. the *training phase* of three emotion classification methods (k-Nearest Neighbours, Support Vector Machine and Random Forests that will be described in detail later);
2. the *classifier accuracy estimation* of the three methods in order to identify the best classification method to be used in the *second experiment*.

Both for training and for the accuracy estimation, we used only the subset *S5*, that is the most inclusive feature subset. In order to train a classifier according to supervising learning approach, we need an input dataset containing rows of features and an output class (e.g. the emotion). The trained classifier provides a model that can be used to predict the emotion corresponding to a set of features, even if the classifier did not use these combinations of features in the training process. According to (Zeng et al., 2009), the most significant classifiers that can be used for

our experiment are *k-Nearest Neighbours* (Cover and Hart, 1967), *Support Vector Machine* (Amari and Wu, 1999) and *Random Forests* (Breiman, 2001). As mentioned above, as final result of the classifier validation, we will select the classification method that provides best performances on emotion recognition accuracy using only the subset *S5*.

Regarding the second part of the first test, to quantify the classification accuracy of the three presented methods, we use the K-Fold Cross Validation Method (K-Fold CRM). More in detail, the k-Fold CRM, after having iterated  $k$  times the process of dividing a database in  $k$  slices, trains a classifier with  $k - 1$  slices. The remaining slices are used as test sets on their respective  $k - 1$  trained classifier to calculate the accuracy and provides as final accuracy value the average of the  $k$  calculated accuracies.

In our case, we impose  $K = 10$ , because this is the number that provides statistical significance to the conducted analysis (Rodriguez et al., 2010). The accuracy estimations obtained with the three investigated methods, k-Nearest Neighbours (with  $k = 1$ ), Support Vector Machine and Random Forests using the subset *S5* to recognize all eight emotions are the following, 85%, 88% and 89%, respectively.

Due to its better performances, we decided to use only the Random forests classifier to conduct the second experiment, that is a full analysis considering all the feature subsets and four different subsets of emotions with numerosity equal to 6, 7, 7 and 8 emotions.

### 3.4 Feature evaluation

The results of the full analysis conducted using the Random Forests classifier (selected after the classifier validation test) are reported in Table ???. As expected, *S4* and *S5* provided better recognition performances with an overall accuracy increment of 6% (in the 6 emotions test) and of 9% (in the 8 emotions test) with respect to that obtained using *S3*. Furthermore, the *Neutral* expression calibration obviously increases the dissimilarity between other emotions.

Comparing the results obtained using the non-differential and differential subsets, in the latter case, it is possible to observe some improvements on the recognition of three particular emotions, *Anger*, *Neutral* and *Sorrow*. The increment of the recognition accuracy of the *Neutral* expression was expected due to the calibration that uses the *Neutral* facial emotion. The increment in the *Anger* and *Sorrow* expressions recognition accuracy was a consequence of the better recognition of the *Neutral* expression since they were often mistaken as *Neutral*. However, we also noticed a decrease of accuracy for the *Disgust* expression recognition using the subset *S4*. In this case, the calibration reduced the *Disgust* dissimilarity in comparison with *Fear*, *Joy*, *Sorrow* and *Surprise*, resulting

in misclassification towards *Surprise* expression.

An interesting result about the classifier performances using subset *S5*, is that it has proved its capacity to exploit the best aspects from the two *S5* subset's components, *S3* and *S4* to improve the emotion recognition accuracy. For example, the classifier used *S3* features to avoid the misclassification of the *Disgust* expression, typical misclassification when using only *S4* features. More in detail, we report in Table 5 and Table 6 the confusion matrices obtained with Random Forests classifier using respectively eight and six (without *Neutral* and *Contemptuous*) emotions for subsets *S3* | *S4* | *S5*. For sake of brevity, we do not report the confusion matrices obtained for the two seven-emotion tests (eight emotions except *Neutral*, eight emotions except *Disgust*), because they provide intermediate results between those achieved for eight and six emotions.

Analysing the literature of the emotion facial recognition systems and comparing them with the obtained results reported in Table ??, we realized that the emotion recognition method based on our proposed features outperformed several alternative methods of feature extraction, presented in Table 3. We compare our method to:

- MPEG-4 FAPS (Pardàs and Bonafonte, 2002), Gabor Wavelets (Bartlett et al., 2003) and geometrical features based on vector of features displacements (Michel and El Kaliouby, 2003) methods with respect to the results obtained by Random Forests classifier using *S3*. These real time methods only classify the six universal facial expressions without using differential features with respect to *Neutral* face with an accuracy of 84%, 84% and 72%, respectively;
- three differential feature methods *Michel et al.* (Michel and El Kaliouby, 2003), *Cohen et al.* (Cohen et al., 2003) and *Wang et al.* (Wang and Yin, 2007) with respect to the results obtained by Random Forests classifier using *S5*. Also these State-of-the-Art (SoA) methods allow the detection of only six universal facial expressions with average accuracies of 73.22%, 88% and 93%, respectively.

To summarize, in Table 3, we report the performance comparison between the aforementioned emotion facial recognition methods considering only the six universal facial expressions emotions (for uniformity of comparison with SoA methods).

Finally, regarding the real-time issue of the emotion recognition system, we calculated that the mean required time (over  $10^3$  tries) to extract our complete proposed set of features (*S5*), once the position of facial landmarks is known, is equal to 1.9 ms. It follows that the working frequencies achievable for sampling and processing, especially when using marker-based landmark locators, are very high and do not compro-

mise the real-time feature of the interaction process.

Table 3: Results using a Random Forests classifier for each dataset composed by a sub-set of features of a sub-set of emotions to classify. \* means without considering contemptuous emotion, \*\* without considering neutral emotion, \*\*\* without considering neutral and contemptuous emotions

No. tested emotions	<i>S1</i> [%]	<i>S2</i> [%]	<i>S3</i> [%]	<i>S4</i> [%]	<i>S5</i> [%]
8	51	76	80	86	89
7*	61	80	84	88	<b>90</b>
7**	60	81	84	<b>90</b>	<b>92</b>
6***	67	87	89	<b>91</b>	<b>94</b>

Table 4: Accuracy comparison of emotion facial recognition methods(not differential or differential features) with six universal facial expressions.

Method	Differential	Accuracy[%]
<i>Michel et al.</i> ( <i>Michel and El Kaliouby, 2003</i> )	No	72
<i>Pardàs et al.</i> ( <i>Pardàs and Bonafonte, 2002</i> )	No	84
<i>Bartlett et al.</i> ( <i>Bartlett et al., 2003</i> )	No	84
<b>Our method <i>S3</i></b>	No	<b>89</b>
<i>Michel et al.</i> ( <i>Michel and El Kaliouby, 2003</i> )	Yes	84
<i>Cohen et al.</i> ( <i>Cohen et al., 2003</i> )	Yes	88
<i>Wang et al.</i> ( <i>Wang and Yin, 2007</i> )	Yes	93
<b>Our method <i>S5</i></b>	Yes	<b>94</b>

## 4 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a versatile and innovative geometric method that extracts facial features inherent to emotions. The proposed method solves the four typical emotion recognition issues and allows a high degree of accuracy on emotion classification, also when compared to complex appearance based methods. Moreover, our method versatility allows the use of different facial landmark localization techniques, both marker-based and marker-less, being a modular post-processing solution. However, it still requires that the face recognition technique presents as output a minimum number of landmarks associated to basic facial features, such as mouth, eyes and eyebrows.

Compared to traditional methods, our method allows, beyond the classification of the six universal facial expressions, the classification of two other emotions: *Contemptuous* and *Neutral*. Therefore, it can be considered as a complete tool that can be incorporated on facial recognition techniques for automatic

Table 5: Confusion matrix with Random Forest using all eight emotions for subsets S3 | S4 | S5.

	Angry	Cont.	Disgust	Fear	Joy	Neutral	Sorrow	Surprise
Angry	<b>76   84   89</b>	09   06   03	02   03   02	00   00   00	00   00   00	07   01   00	05   06   06	00   00   00
Cont.	06   01   02	<b>73   77   82</b>	01   01   01	01   00   00	04   01   00	08   11   08	08   09   09	00   00   00
Disgust	05   03   01	01   00   00	<b>91   89   94</b>	00   01   01	02   04   00	01   01   01	01   02   05	00   01   00
Fear	01   00   00	00   02   01	00   00   00	<b>82   87   87</b>	00   00   00	05   02   01	05   04   05	08   07   07
Joy	02   00   00	01   01   00	01   04   02	00   00   00	<b>95   94   97</b>	01   00   00	01   02   00	00   00   00
Neutral	03   00   00	11   08   06	02   00   00	06   03   01	00   00   00	<b>69   84   87</b>	08   05   03	00   00   00
Sorrow	04   06   02	06   07   06	01   01   01	04   01   03	01   00   00	09   01   03	<b>75   84   85</b>	00   00   00
Surp.	00   00   00	00   00   00	00   00   00	10   07   06	00   00   00	01   00   00	00   00   00	<b>90   93   93</b>

Table 6: Confusion matrix with Random Forest using 6 emotions (without neutral and contemptuous) for subsets S3 | S4 | S5.

	Angry	Disgust	Fear	Joy	Sorrow	Surprise
Angry	<b>86   88   93</b>	03   05   01	00   00   00	00   00   00	10   07   05	00   00   00
Disgust	04   04   02	<b>94   92   96</b>	01   01   01	02   01   00	02   01   01	00   00   00
Fear	01   00   00	01   00   00	<b>86   88   91</b>	00   00   00	07   05   05	07   06   06
Joy	01   01   00	01   04   00	00   00   00	<b>95   96   98</b>	01   01   00	00   00   00
Sorrow	09   06   05	03   01   01	07   02   03	00   00   01	<b>82   91   90</b>	00   00   00
Surprise	00   00   00	00   00   00	08   08   06	00   00   00	00   00   00	<b>92   92   94</b>

and real time emotion classification of facial emotions. As concept proof, we incorporated this tool in a LIFEisGAME (Fernandes et al., 2011) game mode, where the user must match the expression asked by the game. His face is captured and emotion classified in real time. Regarding practical performance, we verified that it is more stable when we apply a neutral face calibration, classifying correctly the emotions expressed. However, it requires that the user knows how to make the expression properly. Problems regarding environment (background and illumination changes) were not addressed. Nevertheless, our method is still restricted to emotion classification of frontal poses, being optimized for static pictures. As future work, we pretend to reduce the landmarks required for emotion classification and to automatize their detection when using unusual face recognition systems. At last but not least, we also pretend to explore sequences of images (including videos) to discover patterns that allow subtle emotions classification, overcoming the limitation of full emotion classification.

## ACKNOWLEDGMENT

This work is supported by PERCRO Laboratory, Scuola Superiore Sant’Anna and Instituto de Telecomunica  es, Funda  o para a Ci  ncia e Tecnologia (SFRH/BD/69878/2010, SFRH/BD/33974/2009), the projects GOLEM (ref:ref.251415, FP7-PEOPLE- 2009-IAPP), LIFEisGAME (ref: UTA-Est/MAI/2009/2009) and VERE (ref:257695). The authors would like to thank to Xenxo Alvarez and Jacqueline Fernandes for their feedback and reviews.

## REFERENCES

- Amari, S. and Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789.
- Asthana, A., Saragih, J., Wagner, M., and Goecke, R. (2009). Evaluating aam fitting methods for facial expression recognition. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE.
- Bartlett, M., Littlewort, G., Fasel, I., and Movellan, J. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW’03. Conference on*, volume 5, pages 53–53. IEEE.
- Bettadapura, V. (2009). Face expression recognition and analysis: The state of the art. *Emotion*, pages 1–27.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Chatterjee, S. and Shi, H. (2010). A novel neuro fuzzy approach to human emotion determination. In *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, pages 282–287. IEEE.
- Cheon, Y. and Kim, D. (2009). Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7):1340 – 1350.
- Cohen, I., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187.
- Cootes, T. and C.J.Taylor (1992). Active shape models - smart snakes. In *British Machine Vision Conference*, pages 266–275. Springer-Verlag.
- Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.
- Cristinacce, D. and Cootes, T. Feature Detection and Tracking with Constrained Local Models. *Biomedical Engineering*, pages 1–10.



- Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.
- Fernandes, T., Miranda, J., Alvarez, X., and Orvalho, V. (2011). LIFEisGAME - An Interactive Serious Game for Teaching Facial Expression Recognition. *Interfaces*, pages 1–2.
- Fischer, R. (2004). Automatic Facial Expression Analysis and Emotional Classification by. *October*.
- Gang, L., Xiao-hua, L., Ji-liu, Z., and Xiao-gang, G. (2009). Geometric feature based facial expression recognition using multiclass support vector machines. In *Granular Computing, 2009, GRC '09. IEEE International Conference on*, pages 318–321.
- Hammal, Z., Couvreur, L., Caplier, A., and Rombaut, M. (2007). Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. *International Journal of Approximate Reasoning*, 46(3):542–567. <ce:title>Special Section: Aggregation Operators</ce:title>.
- Hong, J., Han, M., Song, K., and Chang, F. (2007). A fast learning algorithm for robotic emotion recognition. In *Computational Intelligence in Robotics and Automation, 2007. CIRA 2007. International Symposium on*, pages 25–30. Ieee.
- Jamshidnezhad, A. and Nordin, M. (2012). Challenging of facial expressions classification systems: Survey, critical considerations and direction of future work. *Research Journal of Applied Sciences*, 4.
- Kapoor, A., Qi, Y., and Picard, R. W. (2003). Fully automatic upper facial action recognition. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures, AMFG '03*, pages 195–, Washington, DC, USA. IEEE Computer Society.
- Ko, K. and Sim, K. (2010). Development of a facial emotion recognition method based on combining aam with dbn. In *Cyberworlds (CW), 2010 International Conference on*, pages 87–91. IEEE.
- Kotsia, I., Buciu, I., and Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052–1067.
- Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on*, 16(1):172–187.
- Langner, O., Dotsch, R., Bijlstra, G., and Wigboldus, D. Support material for the article : Presentation and Validation of the Radboud Faces Database ( RaFD ) Mean Validation Data : Caucasian Adult Subset. *Image (Rochester, N.Y.)*.
- Luximon, Y., Ball, R., and Justice, L. (2011). The 3d chinese head and face modeling. *Computer-Aided Design*.
- Michel, P. and El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM.
- Niese, R., Al-Hamadi, A., Farag, A., Neumann, H., and Michaelis, B. (2012). Facial expression recognition based on geometric and optical flow features in colour image sequences. *Computer Vision, IET*, 6(2):79–89.
- Pardàs, M. and Bonafonte, A. (2002). Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing: Image Communication*, 17(9):675–688.
- Rodriguez, J., Perez, A., and Lozano, J. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575.
- Saragih, J., Lucey, S., and Cohn, J. (2011a). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, pages 1–16.
- Saragih, J., Lucey, S., and Cohn, J. (2011b). Real-time avatar animation from a single image. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 117–124. IEEE.
- Seyedarabi, H., Aghagolzadeh, A., and Khanmohammadi, S. (2004). Recognition of six basic facial expressions by feature-points tracking using rbf neural network and fuzzy inference system. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 2, pages 1219–1222 Vol.2.
- Shan, C., Gong, S., and McOwan, P. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816.
- Vetter, T. A Morphable Model For The Synthesis Of 3D Faces f g. *Faces*.
- Wang, J. and Yin, L. (2007). Static topographic modeling for facial expression recognition and analysis. *Computer Vision and Image Understanding*, 108(1-2):19–34.
- Youssif, A. A. A. and Asker, W. A. A. (2011). Automatic facial expression recognition system based on geometric and appearance features. *Computer and Information Science*, pages 115–124.
- Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.
- Zhang, L., Tjondronegoro, D., and Chandran, V. (2012). Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition. *2012 IEEE International Conference on Multimedia and Expo*.





## Appendix C

### VERE Poster: Facial Tracking Systems

## Introduction

For VERE, we explored, tested and developed technologies of extraction and tracking of facial features in real-time. The main goal was to obtain a new approach that is sensitive to subtle and asymmetrical facial movements, is occlusion invariant and, simultaneously, maintains a real time performance using low cost input hardware (e.g. webcams).

Testing these technologies, we concluded that it was necessary to develop a new approach, because the current methods cannot capture subtle and asymmetrical facial features.

## Evaluated Systems

### Optitrack Marker-based facial motion capture

To test this technology, we researched marker configurations (Figure 1) in order to know their sensitivity and sensibility extracting and tracking facial features.

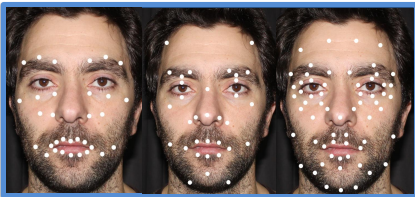


Figure 1: Sample of marker configurations tested: FACS based (left), FaceRobot Autodesk (middle) and The Last of Us (right).

This approach requires specific and high cost hardware to capture and pre and post- process the data. Therefore, we were not able to decode the results.

### Markerless facial motion capture

#### 1. Saragih *et al.* [1] face tracker

We re-implemented Saragih's algorithm [1]. It tracks faces in real time, even with cheaper acquisition hardware (e.g. Webcam) and supports partial face occlusions. Limitations were found when we tried to track asymmetric and subtle movements. Additionally, when applied as feature extraction method we faced a high latency problem, due to high computational cost.

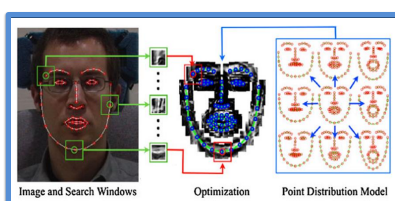


Figure 2: Saragih *et al.* [1] Face tracking algorithm.

#### 2. Microsoft Kinect SDK face tracking

The latency issue presented by Saragih's approach was overcome by the Kinect SDK face tracking system. But, asymmetric and subtle features tracking is still not possible. On the other hand, it requires a Kinect as acquisition hardware and the tracking is highly influenced by lightning changes.

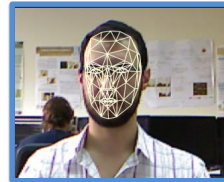


Figure 3: Microsoft Kinect SDK face tracking.

#### 3. Our approach: Displacement maps tracking using Optical Flow

Taking previous results into account, we researched and developed a novel algorithm that allows tracking of basic features in real time (e.g. Saragih *et al.* [1] and Kinect SDK tracker), using cheap acquisition hardware and, additionally, provides subtle and asymmetric features tracking.

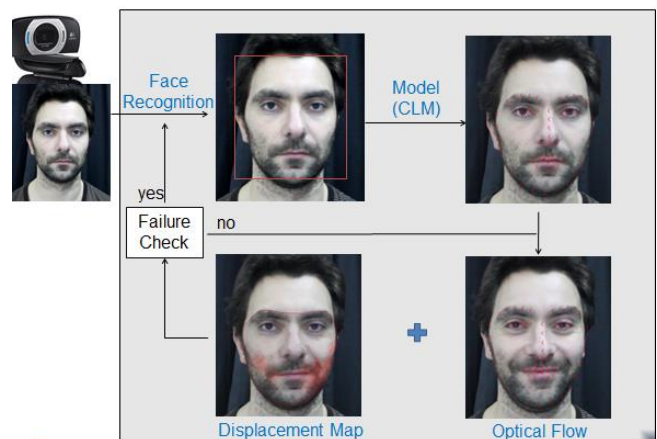


Figure 5: Scheme summarizing our approach.

Test cases using the FdMiee database showed lightning and occlusion invariance; person and expression independence; sensitivity to subtle facial movements. Limitations are still found regarding method stability.

## References

[1] . M. Saragih, S. Lucey and J.F. Cohn, "Face Alignment through Subspace Constrained Mean-shifts," *International Conference of Computer Vision (ICCV)*, September 2009.



## Appendix D

# Facial Expressions Tracking and Recognition: Database Protocols for Systems Validation and Evaluation

---

## Facial Expressions Tracking and Recognition: Database Protocols for Systems Validation and Evaluation

---

**Abstract:** Each human face is unique. It has its own shape, topology, and distinguishing features. As such, developing and testing facial tracking systems are challenging tasks. According to the tracking goals, the researcher needs to collect or combine databases to fit the test and validation procedures specific to that system. However, a database that covers all possible variations of parameters does not exist, increasing researchers' work in acquiring their own data or compiling groups of databases. To address this issue, we propose a methodology for facial data acquisition through definition of fundamental variables, such as subject characteristics, acquisition hardware, and performance parameters. Following this methodology, we also propose two protocols that allow anyone to capture facial behaviors under uncontrolled and real-life situations. As validation, we followed both protocols which lead to creation of two proof of concept databases: FdMiee (Facial database with Multi input, expressions, and environments) and FACIA (Facial Multimodal database driven by emotional induced acting). FdMiee captures facial information under environmental and facial behaviors variations. FACIA is an extension of FdMiee introducing a pipeline to acquire facial behaviors and audio using an emotion-acting method. Therefore, this work eases the creation of databases according to algorithm's requirements and applications, leading to simplified validation and testing processes.

**Keywords:** Computer Vision; Human-Computer Interaction; Performance; Database; Algorithms Validation; Database Protocols.

---

### 1 Introduction

In the field of Computer Vision (CV), there are several existing databases that capture a wide range of facial expressions and behaviors under specific scenarios. The data contained in these databases is usually used for validation and performance tests, as well as training of facial models in CV algorithms [1, 2, 3]. To date, computational works include a limited number of features representing typical facial extraction elements [4, 5, 6]. In fact, there is no single database that integrates a full set of situations: some are dedicated only to expressions, others to lighting conditions, some are just for extracting facial patterns used to define training models, others for emotion classification, etc. This means that the information is split across a variety of databases, making challenging the validation of facial tracking systems under specific situations (e.g. partial face occlusions from hardware or glasses, changes in background, variations in illumination, head pose variations, etc...) or train emotion classifier systems capable of capturing the subtleties of the face using only one database. Aforementioned drawbacks usually lead to systems that only perform accurately in limited environments and facial behaviors [7]. Therefore, every time it is required to design validation and performance tests or training sets, researchers struggle to find databases that fit all system's requirements [7]. As an example, to deploy the face tracking system [8] it was needed the compilation of three different databases. In alternative, researchers define and setup their own procedures to acquire own databases, collecting subjects, defining protocols,

and preparing capture equipment - which are all time-consuming processes. This "database customization" requirement exists since databases require specific features or formats (e.g. high-resolution videos and infra-red pictures) according to CV system's profile and goal. These features and formats contain a wide range of variations in external and facial behavior parameters to simulate real-life situations and provide information that fit the scenario where the system is going to be applied [7].

In this work, we designed two generic protocols and developed a methodology for data acquisition for face recognition systems, as well as for tracking and training of CV algorithms. Our methodology defines each acquisition protocol as the combination of three basic variables: *i)* subject characteristics, *ii)* acquisition hardware and *iii)* performance parameters. These variables are classified as flexible (i.e. can be altered according to system requirements, not influencing protocol guidelines) and fixed (i.e. defined and constrained by the protocol guidelines). The flexible variables are connected to system requirements, and the fixed ones to the information recorded and simulated. As performance variables, we define the following parameters: external (e.g. environment changes in lightning and background) and facial (e.g. variations in facial expressions and their intensity). To test the accuracy and performance of algorithms in facial features tracking or to train face models, used databases need to contain a broad set of external and facial behavior variations. Setting up these variables through our proposed methodology and adopting our protocols eases the process of acquisition of databases with facial information under real-life scenarios and realistic facial behaviors. To validate this process, we followed both protocols and acquired two sample databases. We also analysed the obtained results to establish proof-of-concept.

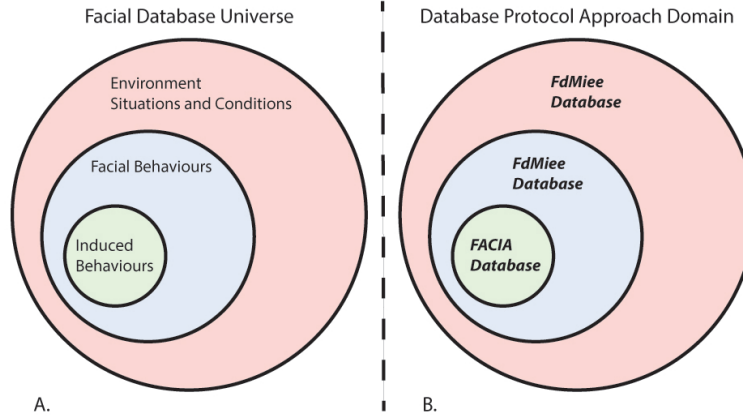
We dubbed the first protocol **Protocol I** that generated *FdMiee* "Facial database with multi input, expressions and environments". Protocol I aims to guide researchers through acquiring data using three capture hardware while varying the performance variable, giving special focus to external parameters variation. As Protocol I's extension, **Protocol II** introduces variations in performance variable regarding facial behaviors. Validation of this protocol generated *FACIA* "Facial Multimodal database driven by emotional induced acting".

Figure 1 represents our overall contribution schematically, regarding the types of data captured in the protocols. It represents the Facial databases' universe through Environment situations and conditions, where we include the group of available facial behaviors, with a small part reserved to introduce behaviors (Figure 1 - A). Taking this scheme into account, we can mirror the domain of our database protocol and represent the contributions of *FdMiee* and *FACIA* diagrammatically (Figure 1 - B).

## 2 Background

To develop guidelines for database acquisition, we researched the literature for methodologies and variance parameters required to test and evaluate CV systems. We analyzed state-of-the-art databases, and classified them into two groups, according to their output format: video and image-based. The most commonly-used video databases are as follows:

- BU-4DFE (3D capture + temporal information): A 3D Dynamic Facial Expression Database [1];

**Figure 1** Summary of database protocols' contributions (B) in facial database universe (A).

- BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database [9];
- MMI Facial Expression Database [2];
- VidTIMIT Audio-Video Database [10];
- Face Video Database of the Max Plank Institute [11].

Comprehensive and well-documented video databases exist, for example [12] and [13]. However, to access them, a very strict license must be procured and a payment provided. BU-4DFE [1] presents a high-resolution 3D dynamic facial expression database. Facial expressions are captured at 25 frames-per-second while performing six basic Ekman's emotions. Each expression sequence contains about 100 frames spread through 101 subjects. More recently, this database was extended to create a 3D spontaneous facial expressions [9]. Another facial expressions database commonly used is the MMI database [2]. It is an ongoing project that holds over 2000 videos and more than 500 images from 50 subjects. Also information of displayed AU's is given with the samples. The VidTIMIT Audio-Video [10] contains video and audio recordings from 43 people reciting 10 short sentences per person. Each person also performs a head rotation sequence per session, which in facial recognition can allow pose independence. Finally, Face Video Database from the Max Planck Institute provides videos of facial action units, used for Face and Object Recognition, though no more information is given [11]. Usage of videos instead of images on the model training allows a better detection of spontaneous and subtle facial movements. However, available databases are limited to standard facial expressions detection [1, 2] or do not explore situations with different lighting levels.

Regarding image-based databases, we came across a comparison study in the table VIII of [3]. This table describes the commonly-used image-based databases for validation of face tracking systems. It also exposes their limitations. As examples of current image-based databases, we analyzed the following databases:

- Yale [14];

- Yale B [15];
- the FERET [16];
- CMU Pose, Illumination and Expression (PIE) [17];
- Oulu Physics [18].

Regarding *Yale* [14] and *Yale B* [15] database, it contains a limited number of grayscale images with well-documented variations on lighting, facial expressions, and pose variations. In contrast, the *FERET* database [16] has a high number of subjects with a complete pose variation. However, no information about lighting is given. Another interesting database is the *CMU PIE* [17] which also tests extreme lighting variations for 68 subjects. These three databases are frequently used for facial recognition, not only for model training but also for validation. Finally, we also highlight the *Oulu Physics* [18] database, since it presents a variation on lighting color (horizon, incandescent, fluorescent, and daylight) on 125 faces.

Based on this research, we concluded that there is a wide range of databases that explore and simulate diverse facial expressions under different environment conditions. However, the available information is spread throughout many databases. In other words, a single database that combines all these facial and environment behaviors and variations providing a complete tool for validation of facial expressions tracking and classification is still non-existent.

In [19], a complete state-of-the-art on emotional databases available nowadays can be found. We searched for a facial expressions database that would simultaneously provide color and depth video (3D data stream) as well as speech information, along with emotional data. Our search criteria, however, were not fulfilled.

The increase of affect recognition CV methods [20] lead to a necessity of databases generation containing spontaneous expressions. To establish how to induce these expressions in participants, we analyzed the review paper on Mood Induction Procedures (MIP's) [21] and investigated which resource materials could be used to enhance and introduce realism in expressed emotions [22]. We concluded that the most commonly-used emotion induction procedure is the Velten method, characterized by a self-referent statement technique. However, the most powerful techniques are combinations of different MIPs, such as Imagination, Movies/Films instructions or Music [21]. Therefore, the technique chosen for our experiment was a combination of the Velten technique with imagination, where we proposed an emotional sentence enacting, similar to the one presented by Martin *et al.* [22].

Some available databases that use similar MIP's induce emotions in the users by asking them to imagine themselves in certain and pre-defined situations [23, 24]. However, the usage of this procedure without complementary material (e.g. sentences) does not guarantee facial expressivity from the user [23, 21]. Since we intended to record speech, we analysed state-of-the-art multimodal databases [19] and found that there was none containing Portuguese speech. Therefore, we decided to explore this potential research avenue.

### 3 Protocol Methodology

Analysing the background and details of facial data acquisition setups, we propose that to create a protocol, three fundamental variables need to be characterized: *subject characteristics*, *acquisition hardware* and *performance parameters* (Table 1).

These variables are classified as being either flexible or fixed, according to their impact on the protocol guidelines. Subject characteristics and acquisition hardware are flexible variables, as they can be changed according to system requirements. For example, use male subjects captured with a high-speed camera or other kind of hardware available, since they do not influence the guidelines of acquisition itself, but only interfere with the acquisition setup. In contrast, fixed variables such as performance parameters, influence guidelines definitions, i.e. different performance parameters require us to take different steps for their simulation and acquisition.

*Subject characteristics* include gender, age, race, and other features that can be extrapolated from the subjects' samples. This variable introduces specific facial behaviors (e.g. cultural variations in emotion expressions) in the database. Regarding, *acquisition hardware*, we enabled the usage of any type of input hardware according to acquisition specifications. Different combinations of these flexible variables can be applied to any of the fixed *performance parameters* guidelines. Performance variables describe the procedures for acquiring the data required for performance tests of CV algorithms. They are split into External and Facial categories, according to what we want to test. External parameters are related to changes in the environment, such as background, lightning, number of persons in a scene (i.e. multi-subject), and occlusions [25, 26, 27]. These variables are almost infinite [28] due to their uncontrolled nature in real-life environments. Facial behaviours should contain facial expressions data triggered by emotions, such as macro, micro, subtle, false, and masked expressions [29, 30, 31] or even speech information. Ekman *et al.* [29] defines six universal emotions: anger, fear, sadness, disgust, surprise and happiness. These universal emotions are expressed in different ways according to a person's mood and intentions. The way they are expressed leads us to an expressions-classification:

**Table 1** Protocol flexible and fixed variables.

Protocol Variables		
Flexible		Fixed
Subjects Characteristics	Acquisition Hardware	Performance Parameters
Gender Age Race (...)	Webcam HD Camera Infra-Red Camera Microsoft Kinect High-Speed Camera (...)	External Parameters: Background Lightning Multi-Subject Occlusions Facial Parameters: Head Rotation Expressions: Macro Micro False Masked Subtle Speech



- **Macro:** These expressions last between half a second and 4 seconds. They often repeat and fit what is being said as well as the speech. Facial expressions of high intensity are usually connected to six universal emotions [29, 30];
- **Micro:** Brief facial expressions (e.g. milliseconds) related to emotion suppression or repression [29, 30];
- **False:** Mirrors an emotion that is deliberately performed, i.e. not being felt [29, 30];
- **Masked:** False expression created to mask a felt macro-expression [29, 30];
- **Subtle:** Expressions of low intensity that occur when a person starts to feel an emotion or shows an emotional response to a certain situation, another person, or surrounding environment. This is usually of low intensity [31].

Facial behaviors generated by speech usually contain a combination of the above expressions [32].

Following this methodology, we developed two protocols. We dubbed the first protocol to generate *FdMiee* **Protocol I**. To validate this protocol, we acquired data from eight subjects with different characteristics. We applied low-resolution, high-resolution, and Infra-red cameras as acquisition hardware variables. As performance parameter variables, we simulated multi-input expressions and environments to test the invariance and accuracy of facial tracking systems exposed to changes, e.g. different lighting conditions, universal-based and speech facial expressions. To validate the results, we executed 360 acquisitions and demonstrated the protocol's potentials to acquire data containing uncontrolled scenarios and facial behaviors. We dubbed the second protocol to create *FACIA* database **Protocol II**. This is an extension of Protocol I's performance parameters variables, introducing induced facial behaviors. To validate the results, we studied the protocol's effectiveness for acquiring multimodal databases of induced facial expressions with speech, color, and depth video (3D data stream) data. To achieve this validation goal, we presented a novel induction method using emotional acting to generate facial behaviors inherent to expressions. We also provided emotional speech in the Portuguese language, since currently there is not any 3D facial database that uses this language. Similar to *FdMiee*, in *FACIA* we created proof-of-concept through an experiment with eighteen participants, in a total of 504 acquisitions.

As a typical protocols' usage example, a research team has available database of 10 female subjects aged between 20-22. They would like to compile a database to test the head rotation tracking accuracy of a CV algorithm using a HD camera. Therefore, they define as *subject characteristics* the female gender and age range. Then, they choose a HD camera as *acquisition hardware* and afterward need to pick the Facial parameter: head rotation as Performance parameter. Finally, they need to follow our validated *FdMiee* protocol.

In summary, to follow the protocols, we first choose the parameters to simulate as fixed Performance variables. This allow us to define the acquisition guidelines. Secondly, we determine the hardware variable and generate an acquisition setup. It is important to note that this variable is flexible, and thus changing this variable will not impact the guidelines. The same is verified using different subject characteristics.

## 4 Protocols and Validation

In this section, we describe in detail the two protocols that follow our proposed methodology. **Protocol I** resulted in the *FdMiee* database that contains facial data from uncontrolled

scenarios. FdMiee focuses essentially on performance variable guidelines of external parameters. The obtained data was recorded with three types of acquisition hardware. As its extension, **Protocol II** focuses on testing and simulation of facial parameters of the performance variables, using Microsoft Kinect as hardware.

#### 4.1 Protocol I

Facial recognition and tracking systems are highly dependent on external conditions (i.e. environment changes) [7]. To reduce this dependency, we developed a protocol based on our proposed methodology, for database creation with changes in terms of external parameters, such as light, background, occlusions, and multi-subject. For facial parameters, we setup guidelines to capture variations in head rotation, as well as universal-based, contempt and speech facial expressions. Table 1 summarizes the performance parameters acquired through this protocol.

##### 4.1.1 Requirements

As protocol requirements, we setup the acquisition hardware and equipment to simulate the selected external and facial parameters.

##### Acquisition Hardware

The chosen acquisition hardware simulates realistic scenarios captured using three types of hardware. To test the protocol guidelines, we chose the following equipment:

- Low-Resolution (LR) camera
- High-Resolution (HR) camera
- Infra-Red (IR) camera

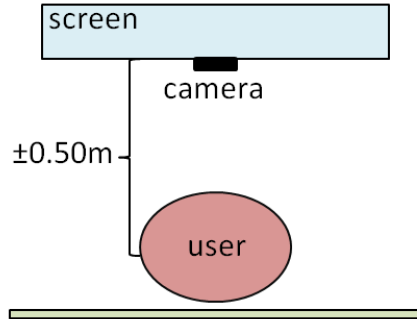
The first two cameras (LR and HR) allow us to study the influence of resolution on face tracking, face recognition, and expression recognition [33]. The IR camera allows us to disregard lighting variations [34, 35, 36] and provides a different kind of information than HR and LR cameras. The hardware used in this protocol should be aligned with one another to ensure future comparison between data acquired with different hardware.

##### Environment-Change Generation Equipment

To generate data with the defined parameters, we stabilize the following environment elements:

**Background** A solid color and static background ease the process of detecting facial features and extracting information from the surrounding environment. The background should ideally be black (or very dark) to prevent interference with the IR camera (black color has lower reflectance compared to lighter colors)

**Lighting** The room must be lit up by homogeneous light, and not produce shadows or glitters in the subject's face. By taking these measure, we ensure that the skin color will have no variations throughout the acquisition process.

**Figure 2** Acquisition setup proposed for Protocol I.

#### 4.1.2 Acquisition Setup

The subject sits in front of the acquisition hardware. This hardware setup is composed of three cameras (LR, HR and IR). The subject's backdrop should be black with some space between them, to have the possibility of moving objects or subjects behind the main scene. This setup is exemplified in Figure 2

#### 4.1.3 Protocol Guidelines

To perform the acquisition, we suggested the presence of two members: one to perform the acquisitions (A) and the other to perform environment variations (B). The subject sits in front of the computer monitor and one of the team members aligns them with the cameras. During the entire acquisition procedure, the subject should remain as still as possible, to avoid producing changes during the various acquisition procedures.

Before starting the experiment, each subject has access to a printed copy of the protocol. This reduces the acquisition time, since the subject already knows what is going to take place during the experiment. Each performance parameter simulated and introduced in the scenario has its own guidelines:

**Control** Team member A takes a photo with the subject in the neutral face.

**Lighting** Team member A takes 3 photos with different exposures (High, Medium, Low). This variable was only acquired in HR camera, because it is the only where it is possible to change the exposure level.

**Background** Team prepare the background to the acquisition.

1. Team member A starts recording;
2. Subject stay still during 5 seconds while team member B performs movement if necessary (only case of dynamic background);
3. Team member A stops recording.

**Multi-Subject** While subject is being record, team member B appear in the scene during 10 seconds.

**Occlusions** For total occlusion, subject will start in the center of the scene and will slowly move to a point out of the scene. For partial occlusions, a photograph is taken with a plain color surface, like a piece of paper covering the following parts of the face:

- Top;
- Left;
- Bottom;
- Right.

**Head Rotation** For each head pose (Yaw, Pitch and Roll) subject performs the movement in both directions while being recorded through the complete movement.

**Universal-Based Facial Expressions, plus Contempt** Subject repeat during 10 seconds the following emotion expressions, starting from the neutral pose to a full pose:

- Joy;
- Anger;
- Surprise;
- Fear;
- Disgust;
- Sadness;
- Contempt.

**Speech Facial Expressions** The subject reads a cartoon or text and is encouraged to express his feelings about it.

#### 4.1.4 Obtained Outputs

This protocol generates the following output data:

- HR and LR Photographies (.jpeg)
- LR camera videos - 15fps (.wmv)
- HR camera videos - 25fps (.mov)
- IR camera videos - 100fps (.avi)

The emotions generated through variation of facial parameters are expected to contain a mixture of macro and micro (i.e. subjects can be repressing and suppressing feelings) as well as false (i.e. subject is making an effort to express certain emotions) and subtle (i.e. when subject cannot generate a high intensity expression) plus speech-based expressions.

### *Data Organization and Nomenclature*

For standardization purposes and further analysis, a folder for each acquisition hardware was created. Inside these folders exist sub-folders for each of the tested performance parameters. The output files were placed in the respective folder with the following template naming convention:

*CaptureModeVolunteer0X\_SimulationName\_take0Y.format*

, where *CaptureMode* is the type of hardware, *X* is the number of the subject, *SimulationName* is the name of the performance parameter acquired and respective information and *Y* is the take's identification number.

#### *4.1.5 FDMie Acquisition & Protocol Validation*

Following the described protocol guidelines, we acquired data from eight volunteers with the following subject characteristics:

**Gender** Male/Female;

**Glasses** With/Without;

**Beard** With different formats/Without.

**Age** 20-35 years

In Figure 3 it is possible to see sample results from some of the performance parameters with the different acquisition hardware.

### *4.2 Protocol II*

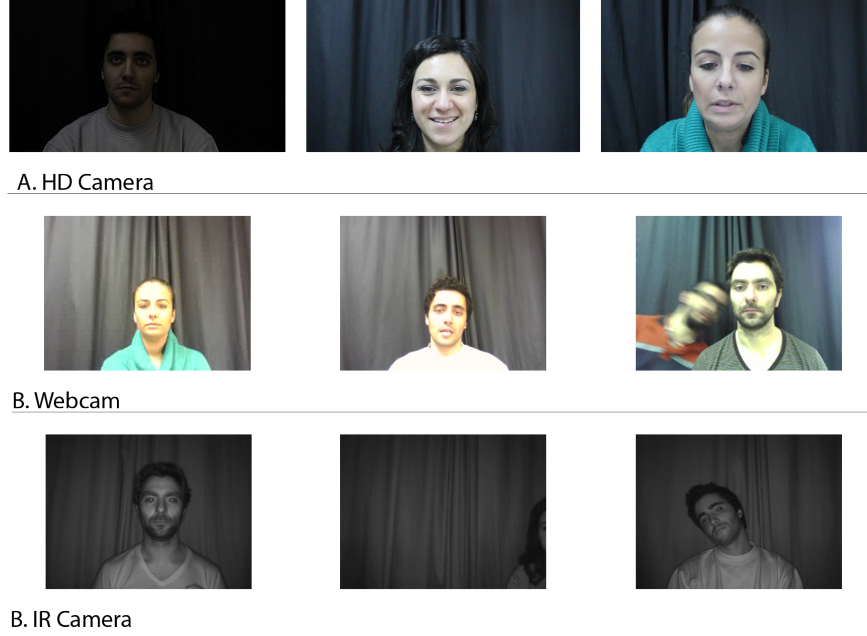
The definition and extraction of induced facial behaviors and speech features inherent to spontaneous expressions is still a challenge for CV systems. To develop and subsequently evaluate a CV algorithm that achieves this goal, we proposed these two protocols to acquire a database containing, simultaneously, spontaneous facial expressions and speech information inherent to induced emotions, such as Ekman's universal emotions [29, 30]. Therefore, in this experiment we focus on the definition of guidelines to capture facial parameters changes in the performance variable 1.

#### *4.2.1 Requirements*

We define two types of requirements: emotion induction method and equipment requirements. Emotion induction method is used as basis to define the protocol guidelines inherent to facial parameters simulation.

#### *The Emotion Induction Method*

The majority of spontaneous facial expressions are generated in real-life situations. To simulate these facial behaviors, we proposed a protocol where the system would ask for emotional acting in order to trigger facial responses from a subject. For this purpose, we combined a Mood Induction Techniques 1 (MIT 1) described by Hesse A.G. *et al.* [21] with mood induction sentences suggested by Pitas I. *et al.* [22]. As an application example,

**Figure 3** FDMiee sample results for HD Camera (A), Webcam (B) and IR Camera (C).

we could have a system that asks for certain user emotions expression through facial or speech features. The user must pronounce certain sentences with a particular tone and facial expression, matching the required emotional state. According to expression classification introduced in Section I, using this method we are able to induce macro, micro, false, masked, and subtle expressions. Macro expressions are implicit, since we ask for expression of the six of the Ekman universal emotions (i.e. anger, fear, sadness, disgust, surprise and happiness). However, since we are in an induced emotions context, subjects can have difficulty engaging in the proposed situation and generating micro, false and masked expressions. Also subtle expressions are triggered because subjects' engaging intensity can be low in the induced sentence or context. As expected, the produced facial expressions depend of subjects' interpretation and how they emerge themselves in the simulated situation.

Our induction approach presents a novel view on emotion acting and their applications, though the domain still remains unexplored in state-of-the-art databases.

We used common persons as subjects, instead of actors, to maintain the natural-ness of real-life scenarios and also achieve a larger diversity of facial behaviors. Actors gain, over time, professional skills that common population cannot reproduce, thus they might introduce features that cannot match the real-world human performance. Some available databases that use MIT 1, try to induce emotions in the users, asking them to imagine themselves in certain general and predefined situations [22, 23, 24]. We also avoided this approach, since suggesting certain situations will not guarantee certain emotion expressions as output by the subject. This is due to the fact that different individuals have different reactions as responses. Therefore, in our protocol, we asked the subjects to imagine and create for themselves some personal mental situation, while they enact the pre-defined

**Figure 4** Example of our video-audio synchronizer (left).

sentence. This aims to ensure an engaging adaptation and natural response from the subject. As mentioned before, the chosen emotions were the six basic Ekman emotions [29], due to their scientific acceptance and applicability in real-world situations. The sentences are pronounced in the European Portuguese language to match the user's mother tongue. This is another contribution of our work, since currently there is not a Multimodal European Portuguese database available.

#### *Equipment and Environment Requirements*

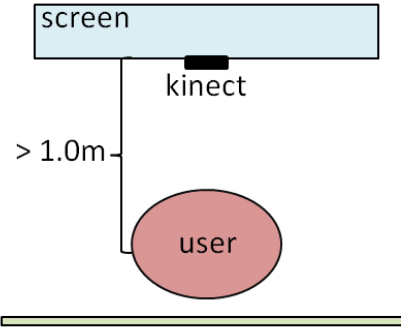
The acquisition setup uses the Microsoft Kinect as acquisition hardware variable. Kinect records 3D data stream as well as speech information. The illumination is not controlled however, as acquisitions were executed during different day periods under uncontrolled lighting conditions. The background is static and white, and there is no sound isolation, since speech signal can be affected by external noise. Sentences are displayed on a screen positioned in front of the subject. To allow further synchronization or re-synchronization, a sound and light emitter is used in the beginning of each recording (see example of Figure 4). For this experiment in our protocol validation, we developed a software that allows simultaneous recording of color and depth video with speech from Microsoft Kinect, in .bin, video, and audio formats. This software includes the Facetracker's Microsoft SDK, and also saves the information retrieved from this algorithm.

#### *4.2.2 Acquisition Setup*

The subject sits in front of the capture hardware. Distance between subject and Microsoft Kinect should be more than 1 meter to enable facial depth capture. A screen displays the sentence that is currently going to be "acted". The subject did not watch the recordings neither observe their own acting, to avoid auto-evaluation or influence their acting performance and expressivity. In FACIA protocol we propose the acquisition setup of Figure 5.

#### *4.2.3 Protocol Guidelines*

Each subject sits in front of the screen and acts out the two sentences per emotion 2 while their voice and face expression are recorded. Per sentence we execute the procedure two

**Figure 5** Acquisition setup proposed by FACIA protocol.**Table 2** FACIA emotion induction method: Sentences pronounced and acted by the subjects.

Emotion	Sentences
Neutral	A jarra está cheia com sumo de laranja
Anger	O quê? Não, não, não! Ouve, eu preciso deste dinheiro! Tu és pago para trabalhar, não é para beberes café.
Disgust	Ah, uma barata! Ew, que nojo!
Fear	Oh meu deus, está alguém em minha casa! Não tenho nada para si, por favor, não me magoe!
Joy	Que bom, estou rico! Ganhei! Que bom, estou tão feliz!
Sadness	A minha vida nunca mais será a mesma. Ele(a) era a minha vida.
Surprise	E tu nunca me tinhas contado isso?! Eu não estava nada à espera!.

times. This ensures the integrity of final results. We suggest a minimum of two members (A and B) in the acquisition team. Before starting the experiment, a protocol describing the experiment is given to the subject.

The experiment starts by a neutral sentence [37] (that can be used as baseline for further experiments).

Therefore, to each sentence of Table 2 the following pipeline is repeated two times:

1. Acquisition team member A says *1,2,3 ... I will record!*.
2. Acquisition team member B uses the light/sound synchronizer.
3. Subject performs the emotion acting.
4. Acquisition team member A stops the recording.

#### 4.2.4 Obtained Outputs

Using our acquisition protocol, we obtained the following data per sentence enacted:

- Video Color - 30fps (.bin);



- Depth image - 30fps (.bin);
- Audio - pcm format - 16000 Hz (.bin).
- Facetracker SDK and Action Units detected (.bin).
- Audio file (.wave).
- Color Video file (.avi).

As explained, regarding facial behaviors we are able to generate data containing macro, micro, false, masked and subtle expressions.

#### *Data organization and Nomenclature*

Similarly to procedure adopted in FdMiee protocol, we predefine how data acquire is going to be organized. To each subject is created a folder called "Volunteer0X", where X is the number associated to the subject. Inside each subject folder are created eight additional folders: one per emotional sentence. Inside of each emotion folder we will have two folders numbered with corresponding sentence, where we will place three data types obtained. Regarding file names, we will use the following template:

*Volunteer0XEmotionSentence0YTake0Z.format*

Where X is the subject number, Y the sentence number and Z the take number.

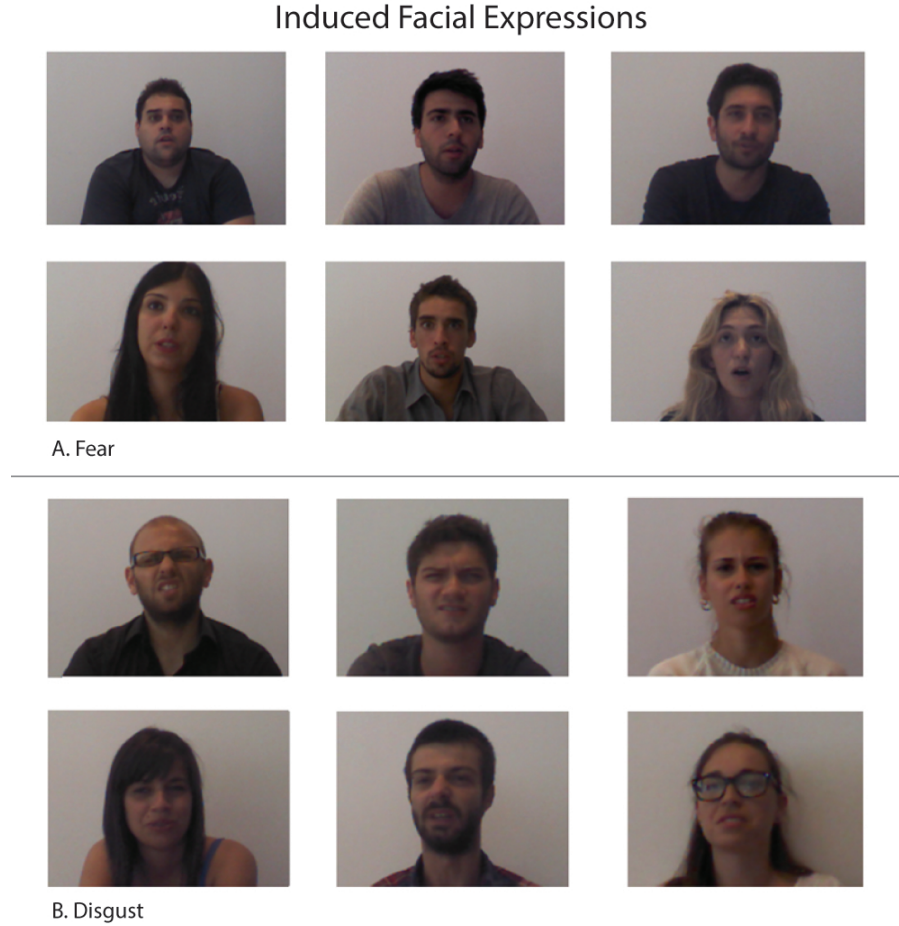
#### *4.2.5 FACIA acquisition & Protocol validation*

To validate Protocol II, we follow it for eighteen subjects, in a total of 130 files per subject (total of 504 acquisitions). As subject characteristics variable we have seven female and eleven male; ages are in a range of 20-35 years old and they were all caucasian. As already explain, we require depth information so a Microsoft Kinect was used as acquisition hardware. As sample of the results acquired during validation we can observed the Figure 6.

## **5 Discussion and Conclusions**

In this paper, we presented a methodology to facilitate the development of two facial data acquisition protocols. Following this methodology, we presented the protocols for simulation and capturing of real-life scenarios and facial behaviors. To validate the protocols, two proof-of-concept databases were created: FdMiee and FACIA. They contain comprehensive information on facial variations inherent to both spontaneous and non-spontaneous facial expressions under a wide range of realistic and uncontrolled situations. Generated databases can be used in a variety of applications, such as CV systems evaluation, testing, and training [7]. They also serve as proof-of-concept. Adopting our methodology and following our protocols reduces the time required for customized database acquisition.

Throughout the protocol creation process, we characterized two groups of variables: flexible variables (subjects' characteristics and capture hardware) and fixed performance variables (external and facial parameters). The first protocol focuses on external parameters' simulation as variation of the fixed performance variable. As an extension, the second

**Figure 6** Sample of results obtained for fear (A) and (B) disgust emotion acting.

protocol provides guidelines to induce and capture real-life facial behaviors as fixed performance variables.

Protocol I allows the acquisition of a facial database containing a large number of fixed parameters' variations (external and facial): lightning, background, multi-subject, occlusions, head rotation, universal-based, and speech facial expressions (Table 1). Lighting variations introduce changes in facial features (e.g. contrast and brightness) [15]. These variations enable us to test how CV systems react to and detect, and how tracking is affected. Static and dynamic variations in the background usually interfere with CV systems' performance while detecting and tracking faces [38]. Therefore, in this protocol, we simulate different background contexts, as well as introduce static and dynamic features in the environment. Similar to background variables, we simulate multi-subject environments, since this situation usually interferes with, and at times, disables CV systems' feature detection [7]. Occlusions generated by glasses or hardware are also common in real-life

scenarios, influencing face recognition and emotion classification accuracy [25, 26, 27]. The increase of Head-Mounted-Displays usage in Virtual Reality applications makes it crucial to test systems invariance while using these variables. Regarding facial behaviors, we reproduced and captured two kinds of facial behaviors - universal-based and speech-based facial expressions. Universal-based Facial Expressions are related to pure emotions [29]. They provide data for emotion recognition systems and enable the testing of systems invariance while subjects' faces change expressions. Speech Facial Expressions, on the other hand, are inherent to all types of expressions [32] (as showed in the image 1) and enable the measuring of systems accuracy and precision. To validate Protocol I, we performed an acquisition on eight subjects with different subject characteristics, leading to the creation of FdMiee database. FdMiee contains facial behaviors under different environment contexts. Hence, this protocol enables the generation of databases that are useful for a wide range of CV systems performance tests.

Protocol II extends the first protocol regarding facial behaviors and performance variables, by introducing induced facial features. To achieve this, we proposed an emotion induction method, where facial expressions were induced through emotional acting. Analysing FACIA generated in the validation process, we verified that facial behaviors inherent to certain emotional acting are indeed different among individuals; i.e. subjects performed different acts to realize identical emotional states. Analysing subjects' facial behaviors, we were able to simulate all types of expressions according to subjects interpretations and engaging in induction sentences. Hence, this protocol provides a large and heterogeneous set of facial behaviors, useful for determining the accuracy of tracking and recognition systems. This was intuitively expected, since expressions inherent to emotional states share some action units [29]. This mixing of expressions can compromise database usage to train a machine learning classifier in pure expressions recognition, increasing classification error. Microsoft Kinect was chosen as the acquisition hardware variable, so that we could record three kinds of data: color, depth (3D facial information) and speech. Introducing depth in the stored data provides valuable information [9]. However, recent studies point out that acquisition rate of Kinect is not sufficient for micro and subtle expressions capturing [29, 30]. This argument explains the poor component of micro and subtle expressions present in FACIA. However, in our methodology we classify this variable as flexible, to ensure that protocol guidelines can be used with other acquisition hardware, i.e. guidelines can be applied with high frame rate cameras and improve the capture of these facial behaviors. The speech recording also allows the Portuguese emotional data collection, opening novel research lines in emotion classification and recognition present in the European Portuguese language speech.

In conclusion, our proposed methodology facilitate the generation of facial data acquisition protocols. This methodology provides a tool for researchers to develop their own facial databases. It also enable performance tests, validation and training processes in CV systems in a wide range of life-like scenarios and facial behaviors, being adaptable to different subject characteristics and acquisition hardware.

## 6 Future Work

Our further work will focus on the following key tasks: First, we aim to enlarge our proof-of-concept databases and, subsequently, perform a statistical validation of the two protocols presented in this paper. Enlarging the databases will provide sufficient data for statistical

validation, using various CV systems. The statistical validation will also provide more measurable information regarding data significance and impact. Second, we aim to devise more parameters for methodology variables to refine the validation process. Third, we aim to introduce a more heterogeneous subject samples, with a wider age range (thus greater presence of wrinkles and facial pigments), skin colors, and make-up. Fourth, we intend to carry out tests with more sophisticated acquisition hardware, such as high-speed cameras. And finally, to increase our work applicability, we intend to extend the fixed variable of performance parameters, providing more guidelines to generate novel situations.

## References

- [1] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.
- [2] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.
- [3] AS Tolba, AH El-Baz, and AA El-Harby. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103, 2006.
- [4] Ashish Kapoor, Yuan Qi, and Rosalind W. Picard. Fully automatic upper facial action recognition. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures, AMFG '03*, pages 195–, Washington, DC, USA, 2003. IEEE Computer Society.
- [5] Yeongjae Cheon and Daijin Kim. Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7):1340 – 1350, 2009.
- [6] Robert Fischer. Automatic Facial Expression Analysis and Emotional Classification by. *October*, 2004.
- [7] D. L. Baggio, S. Emami, D. M. Escriva, K. Ievgen, N. Mahmood, J. Saragih, and R. Shilkrot. *Mastering OpenCV with Practical Computer Vision Projects*. Packt Publishing, Limited, 2012. recommended: advanced OpenCV project support/examples inc. iOS and Android examples.
- [8] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):1–10, 2014.
- [9] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692 – 706, 2014.
- [10] C. Sanderson and B. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. *Advances in Biometrics*, pages 199–208, 2009.
- [11] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

- [12] Rodney Goh, Lihao Liu, Xiaoming Liu, and Tsuhan Chen. The cmu face in action (fia) database. In Wenyi Zhao, Shaogang Gong, and Xiaoou Tang, editors, *Analysis and Modelling of Faces and Gestures*, volume 3723 of *Lecture Notes in Computer Science*, pages 255–263. Springer Berlin Heidelberg, 2005.
- [13] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999.
- [14] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.
- [15] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.
- [16] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The feret evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(10):1090–1104, 2000.
- [17] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.
- [18] E. Marszalec, B. Martinkauppi, M. Soriano, M. Pietika, et al. Physics-based face database for color research. *Journal of Electronic Imaging*, 9(1):32–38, 2000.
- [19] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humane database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pages 488–500. Springer, 2007.
- [20] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.
- [21] Astrid Gerrards-Hesse, Kordelia Spies, and Friedrich W Hesse. Experimental inductions of emotional states and their effectiveness: A review. *British journal of psychology*, 85(1):55–78, 1994.
- [22] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface’05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE, 2006.
- [23] Janneke Wilting, Emiel Krahmer, and Marc Swerts. Real vs. acted emotional speech., 2006.
- [24] Emmett Velten. A laboratory task for induction of mood states. *Behaviour research and therapy*, 6(4):473–482, 1968.

- [25] Shane F. Cotter. Sparse Representation for accurate classification of corrupted and occluded facial expressions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 838–841. IEEE, 2010.
- [26] Ioan Buciuc, Irene Kotsia, and Ioannis Pitas. Facial expression analysis under partial occlusion. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–453. IEEE, 2005.
- [27] Fabrice Bourel, Claude C Chibelushi, and Adrian A Low. Recognition of facial expressions in the presence of occlusion. In *BMVC*, pages 1–10. Citeseer, 2001.
- [28] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [29] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- [30] Clancy W Martin. *The philosophy of deception*. Oxford University Press, 2009.
- [31] Sharon L Buxton, Lorraine MacDonald, and Lynette J Tippett. Impaired recognition of prosody and subtle emotional facial expressions in parkinson’s disease. *Behavioral neuroscience*, 127(2):193, 2013.
- [32] ShashidharG. Koolagudi and K.Sreenivasa Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117, 2012.
- [33] Ying-li Tian. Evaluation of face resolution for expression analysis. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 82–82. IEEE, 2004.
- [34] Lawrence B Wolff, Diego A Socolinsky, and Christopher K Eveland. Quantitative measurement of illumination invariance for face recognition using thermal infrared imagery. In *International Symposium on Optical Science and Technology*, pages 140–151. International Society for Optics and Photonics, 2003.
- [35] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *JIPS*, 5(2):41–68, 2009.
- [36] Saurabh Singh, Aglika Gyaourova, George Bebis, and Ioannis Pavlidis. Infrared and visible image fusion for face recognition. In *Defense and Security*, pages 585–596. International Society for Optics and Photonics, 2004.
- [37] Virginie Beaucousin, Anne Lacheret, Marie-Renée Turbelin, Michel Morel, Bernard Mazoyer, and Nathalie Tzourio-Mazoyer. Fmri study of emotional speech comprehension. *Cerebral cortex*, 17(2):339–352, 2007.
- [38] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K Jain. Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):696–706, 2002.

## Appendix E

### VERE Poster: Real-time facial animation through motion capture

## Introduction

Creating of believable movements in 3D characters using facial mirroring in is a current trend in facial animation. However, fine manual tuning is still required. Mapping algorithms are being developed to reduce this manual requirement. In this work we are focused in optimize the Mapping algorithms. As input we adopt Microsoft Kinect SDK face tracking system for facial features extraction and tracking. Afterwards, taking into account the features tracked, we researched and developed a set of mapping algorithms (Table 1) that transfer properly the motion tracked into the 3D character's rig, automatically producing facial animation.

Direct	Linear interpolation RBF interpolation
Pose - Driven	FACS based [1] Rig poses

Table 1 – Mapping Algorithms.

Each algorithm was evaluated based on:

- Computational performance;
- Believable facial deformations;
- Stability;
- Calibration;
- Pre-defined poses.

## Pipeline Overview

Figure 1 summarizes our pipeline. It is divided in 3 stages:

- 1. Face tracking:** extracts and tracks facial features and, using a Machine learning algorithm, allows 5 Action Units weights classification, per frame.
- 2. Mapping:** transference algorithm that defines how the motion captured triggers the rig.
- 3. Rig:** we implemented 2 versions: one bone based and other boned based with pre-defined poses.

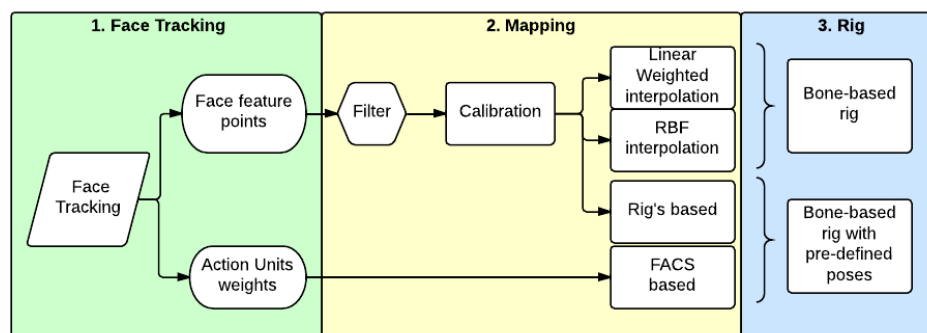


Figure 1: Facial animation motion capture pipeline overview.

## Mapping algorithms

Table 2 summarizes the methods main properties and limitations. To each evaluation factor we make an comparison between methods.

Evaluation Factor	Believable facial deformations	Stability	Calibration	Pre-defined poses	Computational Performance	Error propagation
Linear Weighted interpolation	+	--	-	+	++	-
RBF interpolation	++	-	-	+	+	-
FACS based	+	++	+	-	+++	-
Rig's based	+++	+	--	--	++	+

Table 2 – Mapping Algorithms comparison.

## Results

Figure 2 illustrates the face tracking system used (at left) and the animated 3D character (at right).

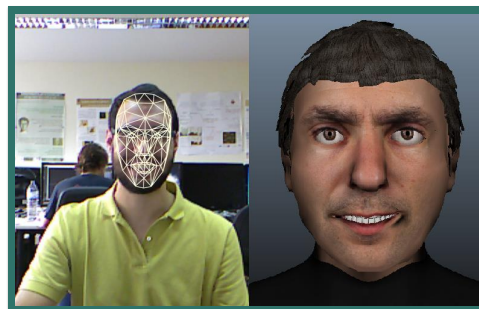


Figure 2:  
Mapping  
algorithm  
example.

Next we intent to improve the rig based method, defining new mathematical transference algorithms that will activate rig's components.





## Appendix F

### Assessing Facial Expressions in Virtual Reality Environments

# Assessing Facial Expressions in Virtual Reality Environments

Catarina Runa Miranda<sup>1</sup> and Verónica Costa Orvalho<sup>1</sup>

<sup>1</sup>*Instituto de Telecomunicações, Universidade do Porto, Portugal*  
*catarina.runa@gmail.com, veronica.orvalho@dcc.fc.up.pt*

**Keywords:** Facial Motion Capture, Emotion and Expressions recognition, Virtual Reality.

**Abstract:** Humans rely on facial expressions to transmit information, like mood and intentions, usually not provided by the verbal communication channels. The recent advances in Virtual Reality (VR) at consumer-level (Oculus VR 2014) created a shift in the way we interact with each other and digital media. Today, we can enter a virtual environment and communicate through a 3D character. Hence, to the reproduction of the users' facial expressions in VR scenarios, we need the on-the-fly animation of the embodied 3D characters. However, current facial animation approaches with Motion Capture (MoCap) are disabled due to persistent partial occlusions produced by the VR headsets. The unique solution available for this occlusion problem is not suitable for consumer-level applications, depending on complex hardware and calibrations. In this work, we propose consumer-level methods for facial MoCap under VR environments. We start by deploying an occlusions-support method for generic facial MoCap systems. Then, we extract facial features to create Random Forests algorithms that accurately estimate emotions and movements in occluded facial regions. Through our novel methods, MoCap approaches are able to track non-occluded facial movements and estimate movements in occluded regions, without additional hardware or tedious calibrations. We deliver and validate solutions to facilitate face-to-face communication through facial expressions in VR environments.

## 1 INTRODUCTION

In the last two decades, we lived a revolution of global digital interactions and communication between humans (Jack and Jack, 2013). We erased geographic barriers and started communicating with each other through phones, computers and, more recently, inside virtual environments using Virtual Reality (VR) headsets. Oculus VR company was the responsible by bringing this hardware to consumer-level making this way of interaction more appealing to common users (Oculus VR 2014). However, VR communications remain a challenge. Human communication strongly rely on a synergistic combination of verbal (e.g. speech) and non-verbal (e.g. facial expressions and gestures) signals between interlocutors (Jack and Jack, 2013). Past communication technologies, like phones and computers, adopted the image stream (e.g. webcams) coupled with speech to transmit both signals creating more realistic and complete experiences (Lang et al., 2012). In VR scenarios, we cannot use image stream since we are interacting with the virtual world embodied in 3D characters (Biocca, 1997; Slater, 2014). As result, the demand for on-the-fly algorithms for 3D characters

animation and interaction is even higher. Ahead of unlocking both communication channels (i.e. verbal and non-verbal), the believable animation of 3D characters using user's movements enhance the three components of the sense of embodiment in VR environments: self-location, agency and body ownership (Biocca, 1997; Kiltner et al., 2012). Even with technological advances in Computer Vision (CV) and Computer Graphics (CG), the reproduction of human's facial expressions as facial animation of 3D characters is still hard to achieve (Pighin and Lewis, 2006). To automatise facial animation, facial Motion Capture (MoCap) has been widely used to trigger animation (Cao et al., 2014; von der Pahlen et al., 2014; Cao et al., 2013; Li et al., 2013; Weise et al., 2011). However, these approaches are not suitable for consumer-level VR applications, requiring or expensive setups (von der Pahlen et al., 2014), manual complex calibrations (Cao et al., 2013; Li et al., 2013; Weise et al., 2011) or do not support the persistent partial occlusion of the face produced by VR headsets (Cao et al., 2014).

To overcome the tracking problem created by persistent partial occlusions, Li *et al.* (Li et al., 2015) proposed a hardware based solution using a RGB-D

camera for capture and strain gauges (i.e. flexible metal foil sensors) attached to VR headset to measure the upper face movements that are occluded. But again, this approach is not suitable for general user. It requires a complex calibration composed by hardware calibration to user and a blendshapes calibration to trigger animation. At the moment, this is the unique on-the-fly facial animation with MoCap solution compatible to VR environments.

**Contributions:** This work delivers and validates consumer-level real-time methods for: (i) facial MoCap method for persistent partial occlusions created by VR headsets and (ii) facial expressions prediction algorithms of occluded face region using movements tracked in non-occluded region. Compared to literature, we reduce user-dependent calibration and hardware requirements, requiring only a common RGB camera for capture. Our methods make current facial MoCap approaches compatible to VR environments and enable the extraction of key facial movements of bottom and upper face regions. The movements tracked and emotions detected can be combined to: trigger on-the-fly facial animation, enabling non-verbal communication in VR scenarios; as input for emotion-based applications, like emotional gaming (e.g. *Left 4 Dead 2* by Valve).

## 2 BACKGROUND

In this section, we aim to study the literature regarding two different topics: (i) facial MoCap solutions for persistent partial occlusions created by VR Head Mounted Displays (HMD) and (ii) partial occlusions impact in facial expressiveness. The first topic presents state of the art facial MoCap solutions to overcome the persistent occlusions' issue. Then, in (ii), we explore how these occlusions restrict face-to-face communication and their impact in face expressiveness. By the end, we search for a connection between occluded and non-occluded facial parts used as guide for methodology definition.

### 2.1 Persistent Partial Occlusions: a today's problem

In literature, we are able to find several promising solutions for real-time automatic facial MoCap (Cao et al., 2014; von der Pahlen et al., 2014; Cao et al., 2013; Li et al., 2013; Weise et al., 2011). However, the arise of VR commercial approaches of consumer-level HMD's (Oculus VR 2014), raised a new issue: the real-time automatic tracking of faces partially occluded by hardware (i.e. persistent partial occlusions

of face) (Slater, 2014). Current MoCap approaches adopt model-based trackers, which produce cumulative errors in presence of persistent partial occlusions (Cao et al., 2014). Therefore, due to the absence of VR devices in mass-market, this issue was almost ignored for years. This resulted in a lack of technological solutions for face-to-face communication for VR environments. Only in 2015, Li *et al.* (Li et al., 2015) highlighted this problem and proposed a hardware based tracking solution. This solution uses an RGB-D camera combined with eight ultra-thin strain gauges placed on the foam liner for surface strain measurements to track upper face movements, occluded by the HMD. The first limitation of this approach is the long initial calibration required to fit the measures to each individual's faces using a training sequence of FACS (Ekman and Friesen, 1978). Also, in subsequent wearings by the same person, a smaller calibration is needed to re-adapt the hardware measures. This training step allows the detection of user's upper and bottom face expressions and activate a blendshape's rig containing the full range of FACS shapes (Ekman and Friesen, 1978). Besides the manipulation complexity, the solution also presents drifts and decrease of accuracy due to variations in pressure distribution from HMD placement and head orientation. As consequence, HMD straps positioning influence eyebrows' movement detection (Li et al., 2015). Li *et al.* solution is currently the only one available to overcome the persistent partial occlusions issue, making this an open research topic in CV algorithms for facial MoCap.

### 2.2 Partial Occlusions and Expressiveness

Everyday, humans' communication use facial expressions and emotions to transmit and enhance information not provided by speech (Lang et al., 2012). Even through technology, we always search for a way to use the non-verbal communication channel. As example, using video stream of our faces; virtual representations, like *emotion smiles*, cartoons or 3D characters with pre-defined facial expressions, etc. Understanding facial expressions and improve their representation in 3D characters is one of the key challenges of CG and plays an important role in digital economy (Jack and Jack, 2013). This role is even more relevant now, with recent advances in VR communications at consumer level (Biocca, 1997). *But how can we use the common solutions of facial animations, like MoCap, if user's face is occluded? Are we able to represent faces using information only from bottom of the face?* To answer these questions, we make a litera-

ture overview regarding several face regions impact in non-verbal communication. The goal is to understand how a partial occlusion of the face affects communication. We also researched for a relationship between occluded and non-occluded facial parts through emotion-based and biomechanics studies. This information was used to build one of this work hypothesis.

In a study about face perception (Fuentes et al., 2013), we concluded that humans have independent shape representations of upper and bottom parts of the face. Similar conclusions are found in emotion perception's literature, where mouth and eyes play different roles (Eisenbarth and Alpers, 2011; Lang et al., 2012; Bombari et al., 2013). In (Eisenbarth and Alpers, 2011; Bombari et al., 2013) it is shown that according to the emotion detected participants used information from eyes, or mouth or both. More precisely, in happy expressions participants used information from the mouth; for sad and angry, from eyes; and to fear and neutral, both mouth and eyes are used. For additional information about non-verbal communication, we forward the reader to (Lang et al., 2012). Taking these statements into account, if we occlude certain region of the face, face-to-face communication is affected and we may not be able to decode expressions properly. Subsequently, the tracking of only certain facial regions, like mouth, is not enough for emotion recognition, for proper communication and to generate believable facial animation of 3D characters.

From the biomechanical point of view, we know that facial muscles work synergistically to create expressions. The muscles interweave with one another, being difficult to decode their boundaries, since their terminal ends are interlaced with other muscles. A detailed research about facial anatomy and biomechanics can be accessed at Chapter 3 of the book *Computer Facial Animation* (Parke and Waters, 1996). Several studies in CG applied the biomechanical approach to create coding systems. These coding systems parameterize human face enabling a faster generation of facial expressions in 3D characters (Ekman and Friesen, 1978; Pandzic and Forchheimer, 2003; Magnenat-Thalmann et al., 1988). Although, they do not provide a clear solution for facial expressions estimation constrained to certain regions of the face. Furthermore, the definition and prediction of facial expressions is even harder when the diversity of facial expressions is considered. Scott McCloud (McCloud, 2006) explains the infinite possibilities of facial expressions combinations (i.e. the way mixing any two of universal emotions can generate a third expression, which, in many cases, is also distinct and recognizable enough to earn its own name) (McCloud, 2006).

Then, analyzing literature, we are able to attain that occlusions generated by VR devices affect communication and using only the information of non-occluded regions is not enough to animate a 3D character. However, biomechanics and facial animation coding systems show a connection between the different facial regions and how diverse and complex is the world of possible expressions. Using these statements, we describe a novel methodology to overcome occlusions problem of facial MoCap and then, to assess facial expressions using non-occluded face information.

### 3 METHODOLOGY

The literature overview of previous section allowed us to formulate the following hypothesis:

*to create a method to estimate facial expressions of upper face and emotions using only bottom face's movements.*

Therefore, we deliver VR consumer-level methods that:

- overcomes the persistent partial occlusions issue in MoCap, making possible the bottom face's movements tracking;
- recognizes universal emotions, plus neutral (Ekman and Friesen, 1975; Jack and Jack, 2013) using bottom face's movements;
- estimates upper face's movements (i.e. eyebrows movements) using information tracked from bottom part of the face.

Figure 1 shows the connection between our VR methods. We start by presenting a method to make generic MoCap systems compatible to persistent partial occlusions produced by VR headsets. Then, applying this algorithm, we are able to track properly the bottom face's features and use them to develop methods that predict the following facial expressions: (i) universal emotions, plus neutral (Ekman and Friesen, 1975; Jack and Jack, 2013) and (ii) eyebrows movements. Combining aforementioned methods, we make possible the MoCap of upper and bottom face movements and estimation of facial emotions under persistent partial occlusions created by VR headsets.

As setup, we suggest the usage of a Head Mounted Camera (HMC) combined with the VR HMD (see Figure 2). At first, we justify the adoption of HMC as capture hardware: When the user is inside the VR environment he is not aware of the space around him.

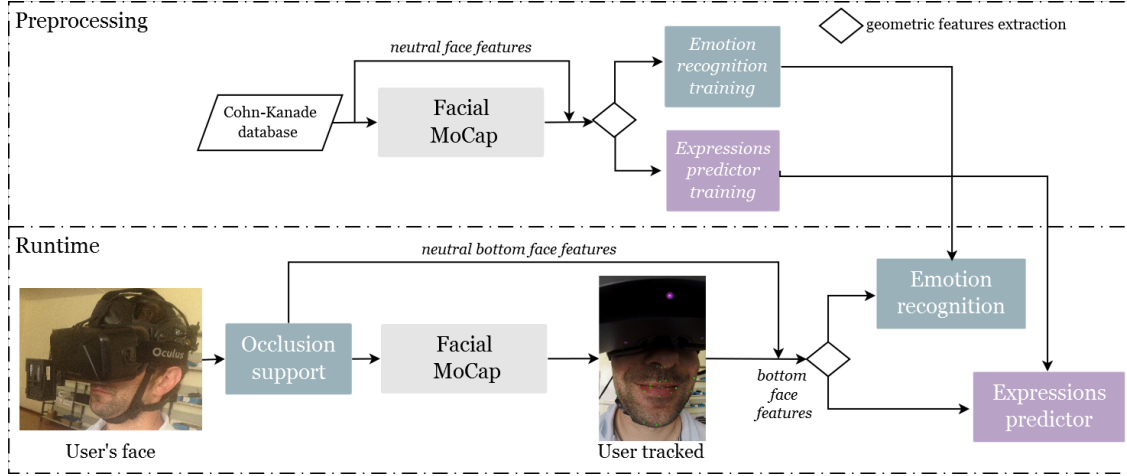


Figure 1: VR methods' framework.

The VR devices precisely substitute the user's sensory input and transform the meaning of their motor outputs with reference to an exactly knowable alternate reality (Slater, 2014). Hence, the user moves and reacts to impulses from VR environment. If we want to capture his face, we have to attach a capture device (i.e. camera) to his body and the device should follow user's movements (see HMC on Figure 2). It is not possible to use a static camera, because the user is not going to be able to place himself in a position proper for capture. A similar setup was also proposed by Li *et al.* (Li *et al.*, 2015), but we removed the strain sensors.

In the next subsections, we provide a complete description of the VR methods.

### 3.1 VR Persistent Partial Occlusions: a novel method

To deploy our occlusion support method for facial MoCap, we used the following statement: we know the kind of occlusion created by HMD, so we know which part of the face is occluded. We also know that MoCap algorithms fail in these situations because they use a face model. When the face is occluded this

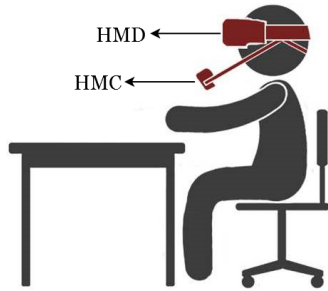


Figure 2: VR setup definition.

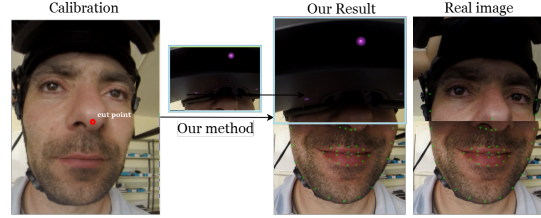


Figure 3: VR method: Persistent partial occlusions. From left to right: calibration image without VR HMD; our method uses cut point (red circle) to cut image an overlay at subsequent images: at left, what facial MoCap method see is a full face and, at right, the real image.

model starts not to fit since there is not a full face being captured. As a solution, we use the knowledge that the region occluded is the upper part of the face to "re-create" the whole face.

Our novel method overlays the upper part of the face captured on a neutral pose during calibration. Firstly, we assume that the higher visible point of the face is the nose and define it as cut point (i.e. this point can be changed to fit the occlusion created by certain HMD). Then, we detect the cut point with the MoCap and we cut the upper part of the calibration image (i.e. frame streamed) from the nose up, and use it to overlay to all the next camera/video frames. Hence, now the occluded part of the face is replaced with a static neutral face. The MoCap system is now able to detect the features in the combined half static/ half expressive face (see Figure 3). We ensure a proper re-creation of a face since we use a HMC that removes the user's head movements, i.e. user's face is in the same position during calibration and next streamed images.

### 3.2 VR Assessing Facial Expressions

During the development of VR facial expressions method, we applied face features and machine learning know-how from our past real-time emotion recognition research (Loconsole et al., 2014). In this novel method, we set the following goals: real-time emotion recognition of universal emotions (Ekman and Friesen, 1975) and upper face expressions prediction under VR scenarios. We aim to track facial expressions ahead of only emotions, in order to get a wide change of facial expressions and better cover and representation of the diversity of faces (McCloud, 1993). In opposition to the emotion classification method (Loconsole et al., 2014), where we needed to reduce the number of features tracked, in VR scenarios we have to maximize the information tracked in the bottom part of the face. Therefore, the feature extraction method should be able to retrieve enough information to allow an accurate prediction of facial expressions by the machine learning algorithm.

As a solution, we propose to use all the features tracked of bottom face region (see Figure 4 blue rectangle) and apply a geometrical features extraction algorithm. This algorithm is defined as the Euclidean distance between neutral face features (stored during calibration step of previous persistent partial occlusions method) and current frame (i.e. instant in time) features. Summarizing, to each feature tracked  $p$  in certain instant  $i$ , we calculate the distance  $D(p_i, p_c)$ :

$$D(p_i, p_c) = \sqrt{\frac{((p_i(x) - p_c(x))^2 + (p_i(y) - p_c(y))^2}{\|p_i - p_c\|}}$$

,where:

$p_i$  is the 2D bottom face feature  $p$  at the instant  $i$  in time;

$p_c$  is the 2D bottom face feature  $p$  of neutral expression captured during calibration;

$\|p_i - p_c\|$  is the norm between  $p_i$  and  $p_c$  in Cartesian space.

Since the occlusion produced varies according to VR headset used, we also created machine learning models to assess facial expressions using the bottom face features information including and excluding nose features. The bottom face features without nose feature can be used by the different kinds of HMD, since the nose region is the one affected by the device size.

To create the machine learning models to predict the emotions and upper face expressions, we used the Cohn-Kanade (CK+) database (Lucey et al., 2010). CK+ database contains posed and spontaneous sequences from 210 participants (i.e. cross-cultural

adults of both genres). Each sequence starts with a neutral expression and proceeds to a peak expression. This sequences are FACS coded and emotion labeled. The transition between neutral and a peak expression allowed us to detect spontaneous expressions and not only pure full expressions.

To implement the algorithms, we adopted a GPU version of Random Forest (Breiman, 2001) of OpenCV (ope, 2014) to generate respective machine learning models for real-time prediction. As facial MoCap testing approach, we deployed the Saragih *et al.* (Saragih et al., 2011) system. (see Figure 4 tracking landmarks in green).

#### 3.2.1 VR Emotion Recognition: novel method

As preprocessing stage, we create the Random Forests model that is used to predict emotions in real-time (Loconsole et al., 2014). To build the model for emotion classification, to each database's sequence we applied the facial MoCap method and extracted bottom face features. Using the first frame of the sequence as neutral expression, to subsequent frames in the sequence, we calculate the distance  $D(p_i, p_c)$ , between bottom face features of current frame and neutral expression's frame. Thus, to train the machine learning model for emotion recognition we used aforementioned geometrical extraction algorithm: distance  $D(p_i, p_c)$  of bottom face's features of each frame. As response value, to each distance calculated, we used respective CK+ emotion label (see Figure 4 blue processes).

As observed in the Figure 2, in runtime, we apply once our occlusions support method and store neutral face features. This step is only execute one time per user. After, in runtime, the adapted facial MoCap system delivers bottom face's movements and distance  $D(p_i, p_c)$  is calculated to each feature  $p$ . The group of distances are used as input in the Random Forests classifier that predicts the user's emotion represented by that distances and respective accuracy's percentage.

#### 3.2.2 VR Facial Expressions Predictor: novel method

To build the upper face expressions model, we also applied the distance of neutral and expression bottom face features as geometric extraction algorithm. However, we have to define the movements that we wanted to predict in order to create specific tags to the training process. For simplicity, we set as upper face expressions the prediction of eyebrows movements, i.e. the detection if eyebrows are going up or down, and the "how much" they are moving compared to

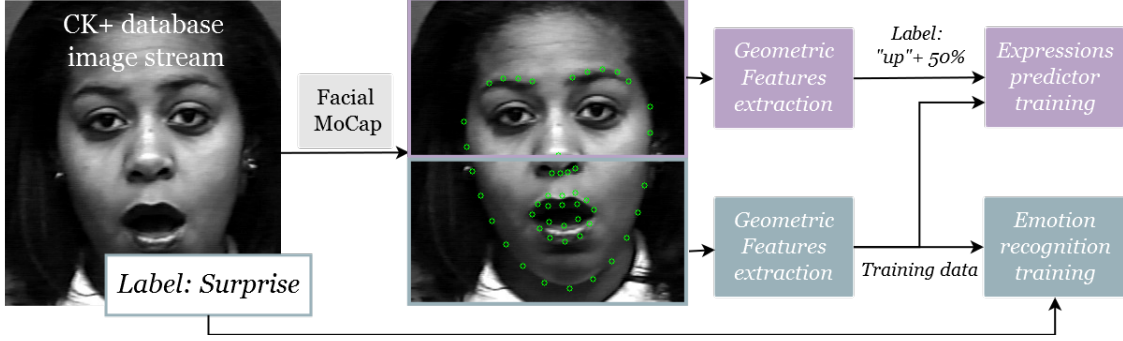


Figure 4: VR methods: Expressions predictor training (purple) and Emotion predictor training (blue) with CK+ database.

a neutral position. This last parameter is measured as a percentage of movement up/down compared to neutral expression. Similarly to assumption made in (Fuentes et al., 2013), we assume symmetry of the eyebrows movements. To define the tags, we calculated the Euclidean distance  $D(p_i, p_c)$  between neutral position of eyebrows and the expression positions in the other frames of the sequence. If the average of the eyebrows features indicated that they are going up, we tagged "up"; the opposite if the eyebrows went down we tag "down" (i.e. we used image coordinate system, so this distance was negative when eyebrows go up and vice-versa). Simultaneously to each frame of the sequence tagged we saved the percentage of movement compared to neutral position (up or down). As result, to each frame of the sequence of each participant in CK+ database we tagged: eyebrows "up" or "down", plus percentage of movement. In Figure 4 with purple processes, the reader can observe an example of method's framework.

At preprocessing stage, we trained two Random Forests models with the same input data: the distances  $D(p_i, p_c)$  between neutral and current bottom face features; but using one of the following response values:

- "up" and percentage of movement, if eyebrows are rising
- "down" and percentage of movement, if eyebrows are descending

, to each frame of each sequence of CK+ database.

Since we are using a GPU approach of the classifier, with high computational performance, to maximize the prediction accuracy of eyebrows movements, we trained two models: one to predict the rise movement and, other, to predict the opposite. In runtime, we apply the defined geometrical features extraction to the bottom face's features tracked by the adapted MoCap. The extracted features are used as input in both Random Forests classifiers, to retrieve one of the predictions:

1. **eyebrows "rising"** and percentage of movement;
2. **eyebrows "descending"** and percentage of movement.

Since we are using two different classifiers, there is a probability of confusion of both models return simultaneously an "up" and "down" movement. As a solution, our method compares the accuracies of prediction from the two classifiers' predictions, and the result delivered is the one with higher accuracy.

## 4 RESULTS AND VALIDATION

In this section, we show the results and statistical validation of the methods proposed. Statistical analysis was performed using R software (R Core Team, 2013).

### 4.1 VR Persistent Partial Occlusions

To test our occlusions method, we applied it to Saragih *et al.* (Saragih et al., 2011) and Cao *et al.* (Cao et al., 2014) MoCap systems (see Figures 5 and 6, respectively). At the Figure 7, we test a generic partial occlusion created by a piece of paper.

As observed in the Figures 5, 6 and 7, our occlusion-support method adapts to MoCap systems making them compatible to persistent partial occlusions. The "paper" test case represented a generic occlusion created by a random VR device. As conclusion, our method is not only adaptable to MoCap, but it could be also used to generic partial occlusions created by different VR HMD's.

### 4.2 VR Assessing Facial Expressions

We divided the validation of our prediction methods in two steps: (i) statistical validation and (ii) visual validation.





Table 1: k-Fold CRM Accuracy comparison to scenario (i) and to the scenario (ii). Results in percentage (%).

Emotions	k-Fold Accuracy (%)	95% Confidence Interval
Six (Ekman and Friesen, 1978)	64.80	[61.72;67.79]
Four (Jack and Jack, 2013)	69.07	[65.59;72.40]

features of the face allow a weak recognition of face emotions, resulting in accuracies lower than 70%.

More in detail, we report in the Tables 2 and 3, a statistical analysis of each emotion recognition obtained with Random Forests classifier to scenario (i) and (ii), respectively.

Both statistical analysis resulted in a p-value lower than  $2.2 \times 10^{-16}$  to a significance level of 5%, which validates our method’s hypothesis: classifying the six/four universal emotions using bottom of face features tracking. Specifically, to scenario (i) at the Table 2, we observe an overall low sensitivity to emotions classified (with exceptions to Joy/Happy and Neutral). The opposite is observed to specificity. This indicates that the method does not have high accuracy to detect a certain class, however, does not predict incorrectly. The predictive values weighted using information about the class prevalence in population, show an overall increase of accuracy for true positive and maintain to negative. Therefore, as example to Surprise, despite our classifier only being able to positively identify surprise in 59.40% of the time there is a 71.82% chance that, when it does, such classification is correct. Looking to Table 3, compared to previous results of scenario (i) at Table 2, we observe an increase of sensitivity, while maintaining an high accuracy of specificity. In general, the same is observed in positive and negative predictive values. This is expected, since decreasing the number of classes of emotions will decrease the degree of confusion that lead to a better split between classes, resulting in a better emotion recognition method. These results confirm the statement of Background section, i.e. bottom face features provide incomplete information about face expression of emotions. Though, our method presents better performance when four universal emotions (Jack and Jack, 2013) are classified.

#### 4.2.2 VR Facial Expressions Predictor

To analyze and validate the VR facial expressions predictor, we executed the k-Fold cross-validation to the classifier eyebrows ”rising” and to classifier eyebrows ”descending”. Taking into account the variance of nose tracking with the type of HMD used, we propose to study the influence of tracking these features (subset *S1*) and not tracking the nose features (subset *S2*) in the prediction of eyebrows’ movements. Average K-Fold CRM accuracies and respective confi-

dence intervals can be accessed in the Table 4.

In the Table 4, we observe a small decrease of accuracy when the nose features tracking is removed. Although, the confidence intervals show that this decrease is only significant in eyebrows ”up” detection. Our method allows an high performance of eyebrows ”up” estimation (at least, 85%) compared to eyebrows ”down” estimation (at least, 66%). The different results arise from the fact that we are using an emotion database for training, where there is more data describing the ”rising” movement than the opposite (i.e. only anger and sadness emotions usually present this facial expression behavior (Ekman and Friesen, 1978)).

Similarly to emotion recognition method, we present the statistical analysis of sensitivity/specificity and positive/negative predictive values to both eyebrows movements using the subsets *S1* and *S2*.

Table 5: Eyebrow Up prediction - Statistical Analysis to subsets *S1*. Results in percentage (%).

Eyebrows Up	<i>S1</i>	<i>S2</i>
Sensitivity	97.34	96.27
Specificity	71.79	59.18
Positive pred.	92.04	87.65
Negative pred.	92.31	84.06

Both p-values of further analysis are lower than the significance level (i.e. p-value equal to  $2.2 \times 10^{-16} < 0.05$ ). Therefore, both methods are suitable for eyebrows movement estimation using bottom face’s movements. Table 4 shows that the method is able to classify the eyebrows ”up” movement accurately, with exception for specificity using the subset *S2*. So, the removal of nose features tracking leads, essentially, to a decrease in accuracy of the classifier in not giving incorrect predictions. However, when we take in to account the prevalence of the class in population, the overall accuracy of prediction to both positive and negative values increase, presenting values above 84.04%.

Table 6 contains the statistical analysis to the prediction of eyebrows ”descending” movement with (*S1*) and without (*S2*) nose features tracking.

Observing the Table 6, we observe that our method predicts correctly the ”descending” movements of the eyebrows, at least, 73.18% of the time and does not predict incorrectly this movements in

Table 2: Statistical Analysis of scenario (i) - Results in percentage (%).

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Sensitivity	53.15	39.44	26.09	81.29	12.70	59.40	90.80
Specificity	86.55	97.70	95.84	95.17	99.13	96.35	85.39
Positive pred.	40.21	57.14	39.34	75.90	50.00	71.82	75.51
Negative pred.	91.56	95.40	92.62	96.45	94.31	93.81	94.92

Table 3: Statistical Analysis of scenario (ii) - Results in percentage (%).

	Anger	Joy	Sadness	Surprise	Neutral
Sensitivity	75.50	77.85	13.80	68.75	80.09
Specificity	76.16	95.14	99.07	98.39	91.34
Positive pred.	45.06	81.46	66.67	88.51	80.44
Negative pred.	92.31	94.00	89.52	94.59	91.16

Table 6: Eyebrow Down prediction - Statistical Analysis to subsets  $S1$ . Results in percentage (%).

Eyebrows Down	$S1$	$S2$
Sensitivity	77.13	73.18
Specificity	62.73	63.97
Positive pred.	71.57	72.09
Negative pred.	69.28	65.23

at least, 63.97% of the time. The lower values are obtained to the subset  $S2$ , however, the differences between subsets performance are not significant. Similar behavior is beheld taking into account the prevalence of the class in the population. The positive/negative predictive values are not significantly different between sensitivity/specificity. As expected by previous k-Fold CRM results, prediction of the "descending" movement presents lower performance compared to prediction of the opposite movement. Again, this result occurred due to the low prevalence of the "down" class in population. This statement is confirmed by the lower influence shown in positive and negative predictive values when compared to sensitivity and specificity, respectively.

Summarizing, our methods of facial expressions prediction are suitable for the estimation of eyebrows movements using features from the bottom of the face, specially in estimation of the "rising" movement. This conclusion corroborates the hypothesis of this work: our results traduce a connection between bottom and upper face behaviors.

#### 4.2.3 VR Assessing Facial Expressions: Visual Validation

Applying the methods to videos where the participants expressed emotions (Ekman and Friesen, 1975), we are able to check visually the performance of the methods: occlusions support, emotion recognition and expressions prediction. We chose a non-VR

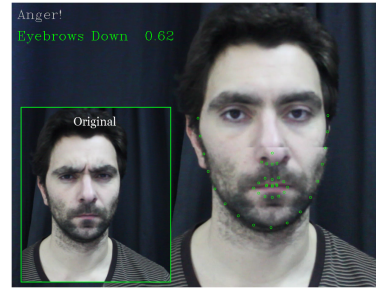


Figure 8: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (green). Check that our emotion and prediction match original image eyebrows movements (green box).

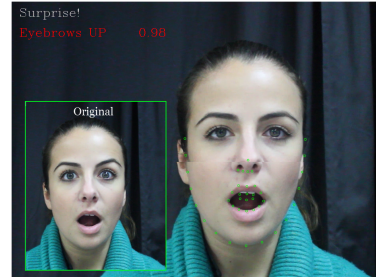


Figure 9: VR Assessing Facial Expressions: Emotion Recognition result (blue) and Expression Predictor result (red). Check that our emotion and prediction match original image eyebrows movements (green box).

scenario in order to verify if the upper face movements and emotions predicted (using only bottom face's movements) match the original facial expressions. Results can be observed in the Figures 8, 9, 10 and 11.

Looking throughout the Figures, we verify that our occlusion method is able to "re-create" the face even not using a HMC. Regarding emotion recognition using only the facial features (green dots), in the Figure 8, 9 and 10, we show three examples of correct classification. Figure 11 presents an example of a wrong emotion recognition. The classifier returned

Table 4: k-Fold CRM Accuracy comparison facial expressions assessed (Eyebrows Up or Down) with subset *S1* and *S2*. Results in percentage (%).

Eyebrows movements	k-Fold Accuracy(%)	95% Confidence Interval
Up <i>S1</i>	91.47	[89.76;92.98]
Up <i>S2</i>	87.02	[84.97;88.89]
Down <i>S1</i>	70.63	[67.99;73.18]
Down <i>S2</i>	69.13	[66.40;71.76]

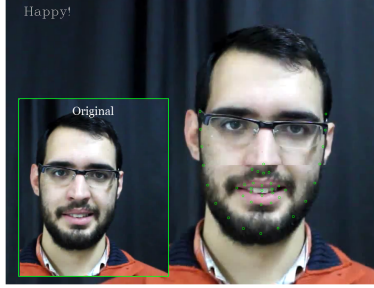


Figure 10: VR Assessing Facial Expressions: Correct Emotion Recognition result (blue) and no Expression Predictor result, since there is not movement. Check original image in green box.

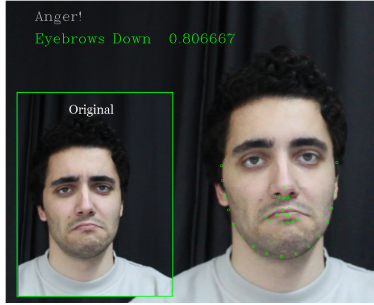


Figure 11: VR Assessing Facial Expressions: Incorrect Emotion Recognition result (blue) and Expression Predictor result. Check original image to see that Expression Predictor is correct (green box).

Anger when the user's emotion label of the video was Sad. This confusion is predicted since the bottom features inherent to Anger and Sad emotions are identical (Ekman and Friesen, 1975).

Regarding the facial expressions prediction method, in the Figures 8 and 11 we observed that the algorithm correctly estimates eyebrows "down", which is confirmed by the original images. The same is detected in the Figure 9 for eyebrows "up" predictor. Moreover, in the Figure 10, comparing eyebrows of image analyzed and original image, we observe no movement, which traduced in a correct no estimation of movement from both predictors.

## 5 CONCLUSIONS

This work delivers VR consumer-level methods to achieve the three goals: make MoCap systems compatible to persistent partial occlusions, real-time recognition of universal emotions and real-time prediction of upper face movements using bottom face features tracking. Combining the three methods deployed, we are able to track in real-time facial expressions from non-occluded and occluded facial regions. The development of these methods lead to improvement in the three components of sense of embodiment, i.e. enhances the sense of self-location, agency and body ownership within the VR environments (Kiltner et al., 2012).

Analyzing the results, we conclude that the three goals proposed where achieved. We deliver a method to make MoCap systems able to track bottom face features under generic partial occlusions created by different HMD's. Note, we do not deliver a method that is able to overcome generic and unpredicted facial occlusions, since we require the knowledge of the area occluded. Then, using these facial features, we were able to define methodologies to real-time recognition of four universal emotions (Jack and Jack, 2013) with an accuracy of 69.07% and prediction of facial movements in the occluded regions, i.e. eyebrows "rising" with accuracy of 91.47% and "descending" with an accuracy of 70.63%. The results obtained with the facial expressions prediction method confirmed our method's hypothesis. Therefore, besides bottom features of the face being not enough to describe the six emotions of Ekman and Friesen (Ekman and Friesen, 1975), our predictor of facial expression decode a connection between bottom face and upper face features. As explained in methodology, the combination of both emotion and expressions tracked/predicted make us able to access a wide range of facial expressions enabling us to represent the diversity of faces (McCloud, 1993). This conclusion opens new lines of research to predict more complex movements of the face, even when we are not able to track them using CV algorithms. Furthermore, our methods outputs enable the real-time animation of 3D characters, since we deliver information of facial features combined to emotions, suitable to activate different types of rigs.

Ahead of 3D characters animation, our methods are suitable for emotion-based applications, like affective virtual environments, advertising or emotional gaming.

As future work, we aim to define a transfer algorithm and use movements and emotions estimated to trigger facial animation. Furthermore, we intend to study how the estimation of more facial behaviors information (e.g. forehead and eye movements) and combination of speech data can improve the animation and user embodiment in VR environments.

## ACKNOWLEDGEMENTS

## REFERENCES

- (2014). OpenCV.
- Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, 3(2):0–0.
- Bombari, D., Schmid, P. C., Schmid Mast, M., Birri, S., Mast, F. W., and Lobmaier, J. S. (2013). Emotion recognition: The role of featural and configural face information. *The Quarterly Journal of Experimental Psychology*, 66(12):2426–2442.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- Cao, C., Hou, Q., and Zhou, K. (2014). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43.
- Cao, C., Weng, Y., Lin, S., and Zhou, K. (2013). 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41.
- Eisenbarth, H. and Alpers, G. W. (2011). Happy mouth and sad eyes: scanning emotional facial expressions. *Emotion*, 11(4):860.
- Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.
- Ekman, P. and Friesen, W. V. (1975). Unmasking the face: A guide to recognizing emotions from facial cues.
- Fuentes, C. T., Runa, C., Blanco, X. A., Orvalho, V., and Haggard, P. (2013). Does my face fit?: A face image task reveals structure and distortions of facial feature representation. *PLoS one*, 8(10):e76805.
- Jack, R. E. and Jack, R. E. (2013). Culture and facial expressions of emotion Culture and facial expressions of emotion. *Visual Cognition*, 00(00):1–39.
- Kiltner, K., Groten, R., and Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387.
- Lang, C., Wachsmuth, S., Hanheide, M., and Wersing, H. (2012). Facial communicative signals. *International Journal of Social Robotics*, 4(3):249–262.
- Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P.-L., Nicholls, A., and Ma, C. (2015). Facial performance sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4).
- Li, H., Yu, J., Ye, Y., and Bregler, C. (2013). Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4).
- Loconsole, C., Runa Miranda, C., Augusto, G., Frisoli, G., and Costa Orvalho, v. (2014). Real-time emotion recognition: a novel method for geometrical facial features extraction. *9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2014)*, 01:378–385.
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010 IEEE Computer Society Conference on, pages 94–101. IEEE.
- Magnenat-Thalmann, N., Primeau, E., and Thalmann, D. (1988). Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297.
- McCloud, S. (1993). Understanding comics: The invisible art. Northampton, Mass.
- McCloud, S. (2006). *Making Comics: Storytelling Secrets Of Comics, Manga And Graphic Novels* Author: Scott McCloud, Publisher: William Morrow. William Morrow Paperbacks.
- Pandzic, I. S. and Forchheimer, R. (2003). *MPEG-4 facial animation: the standard, implementation and applications*. Wiley. com.
- Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45.
- Parke, F. I. and Waters, K. (1996). *Computer facial animation*, volume 289. AK Peters Wellesley.
- Pighin, F. and Lewis, J. (2006). Performance-driven facial animation. In *ACM SIGGRAPH*.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rodriguez, J., Perez, A., and Lozano, J. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575.
- Saragih, J. M., Lucey, S., and Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215.
- Slater, M. (2014). Grand challenges in virtual environments. *Frontiers in Robotics and AI*, 1:3.
- von der Pahlen, J., Jimenez, J., Danvoye, E., Debevec, P., Fyffe, G., and Alexander, O. (2014). Digital ira and beyond: creating real-time photoreal digital actors. In *ACM SIGGRAPH 2014 Courses*, page 1. ACM.

Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011).  
Realtime performance-based facial animation. *ACM  
Transactions on Graphics (TOG)*, 30(4):77.



## Appendix G

Facial emotions as a metric of 21st Century competencies: a draft protocol for acquiring and classifying facial emotion data for learning analytics

# **Facial emotions as a metric of 21st Century competencies: a draft protocol for acquiring and classifying facial emotion data for learning analytics**

**Stephen Haggard, Independent Consultant, UK**

**Veronica Orvalho, Universidade do Porto, Instituto de Telecomunicações, Portugal**

**Catarina Runa, Universidade do Porto, Instituto de Telecomunicações, Portugal**

Emotional states are thought to play an influential role in the exercise of 21<sup>st</sup> Century competencies and in education processes generally, but up to now this possibility has not been capable of rigorous investigation due to lack of appropriate non-intrusive capture technology, and problems of classifying emotional phenomena. The value of emotional data for a learning analytics of 21<sup>st</sup> Century skills is located in its potential to offer a broad and versatile metric that may contribute as a component to several identified competencies. A tool, is described for capturing emotional data unobtrusively from learners, using facial emotions recognition, and classifying them on the Ekman 5-dimension scale of emotions. The protocol is presented for its forthcoming application in a school to capture emotional traces at cohort and individual specific, classify the expressions, and attempt to correlate these emotion values to learning events. The capture and classification technique is proven in other settings, although limitations around capture and interpretation are expected for an education application.

Keywords: Facial emotions, non-cognitive competencies, 21<sup>st</sup> Century skills, learning analytics, emotional data, classification of emotions

## **1. Emotion in the 21<sup>st</sup> Century Skills Canon**

Emotion features in taxonomies of 21st Century Skills as specific component in interpersonal skills, and as an implicit element in personal effectiveness (“[PDF]New Vision for Education Report - Weforum.org - World Economic Forum,” n.d.), and sometimes not at all (OECD 2013). Examination of the role of emotion in the non-cognitive skillset by National Academy of Sciences focussed on the difficulties of assessment, with Stephen Fiore noting that emotions are context-dependent and that fairness and bias of measurement are not yet understood. (Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, and National Research Council 2011) On first view, then, emotion does not qualify as a primary element in any proposed frameworks for 21st Century Skills.

However, in the discipline of learning analytics, measures of emotion may usefully inform the proposed dimensions for the 21st Century Skills canon, such as grit, leadership, intuition or cross-cultural working, whose measurement, it has been argued, will require multifaceted approaches (Shechtman et al. 2013) and in which emotion is arguably a common or underlying factor (Antonio 1995). Emotional states are typically manifested through readily observable behaviours and are therefore easily accessible for recording, while their description and measurement has been the subject of extensive research, and a range of data capture methods are available.



Additionally, the role of emotion in the education process is attested from several perspectives such as education psychology research, and classroom practice. We briefly characterise the educational literature on emotion below.

Education psychologist Pekrun has most recently reviewed research in the “nascent field” of emotions in education and offers a secure conclusion that learning outcomes are determined by learner’s affective states such as interest, anxiety, shame, curiosity, confusion and boredom (Pekrun and Linnenbrink-Garcia 2014). Methodologies for observing and measuring emotions in classroom settings are, on the other hand, diverse and possibly hard to operationalise, including self-reporting, observation, nervous system arousal, neuroscientific techniques including neuro-imaging, and situated approaches. (Immordino-Yang and Christodoulou 2013). This fragmentary approach to evidencing emotion leads Pekrun to assert that techniques should be multidimensional, and to exhort researchers to deliver “evidence-based recommendations for educational practice”.

Classroom practice typically engages with emotion as a potentially problematic manifestation in learners, to be addressed through behavioural training programmes. Famington’s review of non-cognitive competencies as an underlying success factor in formal attainment (Famington et al. 2012) confirms that the student emotional datum, potentially the target of interventions in social and emotional learning (SEL), is clearly important, but detects “little coherence to the broad array of research findings”

Meanwhile, critical accounts identify “the new orthodoxy of emotion”, and even warn that observing emotional factor risks of “pathologising” learners through misinterpretation of socially embedded differences. (Gillies 2011).

Pedagogy is, then, interested in the topic of emotion, and it appears to be a versatile and broadly applicable data class in learning settings. However we lack a convincing theoretical account of emotion as a variable in shaping learning outcomes.

Some models for emotion as a determinant of adaptive behaviours can be found in the environment of videogames. For this environment, the emotions, behaviours and physiological state of players are routinely engaged applied to increase game immersivity (Ventura, Shute, and Small 2014). This approach can be called Emotional Gaming. As an example, in *Left 4 Dead 2*, from game-maker Valve, user levels of the emotions of stress and excitement are recorded using brainwave-reading headsets that monitor EEG signals. The game adjusts and adapts the difficulty and intensity of visual inputs in real-time according to how excited or stressed the players are at that moment and enhances player persistence (DiCerbo 2014). The model for persistence is that the level of struggle to overcome game obstacles is customised to address the player’s state of emotion, leading to increased enjoyment and reduced frustration.

Whether analogous mechanisms can be operationalised in learning settings, with the measuring of emotion potentially contributing through an analytics process to some kind of learning optimisation, is moot. But this question is now open for study, and possibly for

insights, thanks to the availability of a cheap, non-intrusive and proven and high fidelity method for recording emotional data. We present this method below.

## **2. Emotion Classification Techniques**

On the generally accepted scale for classifying and measuring the six universal Ekman emotions (i.e. joy, sorrow, surprise, fear, disgust and anger) (Ekman and Friesen 2003), there are several established methods for data capture. Of these, Computer Vision (CV) with webcams for recording facial movements has general acceptance among users for its non-intrusive nature compared to Brain Computer Interaction technologies (e.g. fMRI, EEG). However, a limitation of traditional CV approaches to date is that they fail to integrate in a single method the following features: real-time performance; capability of recognition multiple universal emotions; recognition of emotions without neutral face calibration; reduction of user manual interventions (Jamshidnezhad and Nordin 2012). Overcoming this limitation allows the development of consumer-level applications, such as educational tools and mood-based games, which can detect and classify in real-time facial emotions. In 2014 we presented and validated a novel geometrical method for emotion classification (Loconsole et al. 2014), which overcomes the aforementioned limitation, as it is able to recognize the six universal emotions (Ekman and Friesen 2003) with 94% accuracy.

Because we now have a feasible emotion measurement technology, we can attempt to establish baseline measures for the emotional components of some non-cognitive skills. We believe this can be a practical and theoretical contribution to validation of pedagogic models required for a learning analytics of 21<sup>st</sup> Century competencies, and will help researchers to evaluate the potential of emotions as a data class that might contribute to a multifaceted measurement approach in a learning analytics for the 21<sup>st</sup> Century Skillset.

In the remainder of this paper we present our protocols for a forthcoming education implementation of CV facial emotion recognition, describe the experimental procedure, and offer some preliminary discussion around issues in implementing this data class in learning analytics. We hope eventually to be able to theorise and explore the relationship between emotions as accessed by universal facial expressions (Jamshidnezhad and Nordin 2012) and the 21st Learning Competencies. However, our current more modest aim is a technical proof in learning contexts for our methodology for tracking the six basic emotions; and a first excursion into the challenges of mapping such observations to the non-cognitive competencies.

## **3. Implementing the Geometrical Facial Emotions Classifier in Education**

Our current geometrical emotion classifier technology, which runs as a standalone application and has been validated and tested in several real-life scenarios (Alves et al. 2013) is to be adapted and applied to capture of learners faces during a range of pre-defined standard formal learning contexts (e.g. a maths exercise, literacy exercise). In Autumn 2015 we are conducting a school-based technology trial geared at resolving hardware, data collection, feasibility and acceptance issues, and which will also yield a provisional database

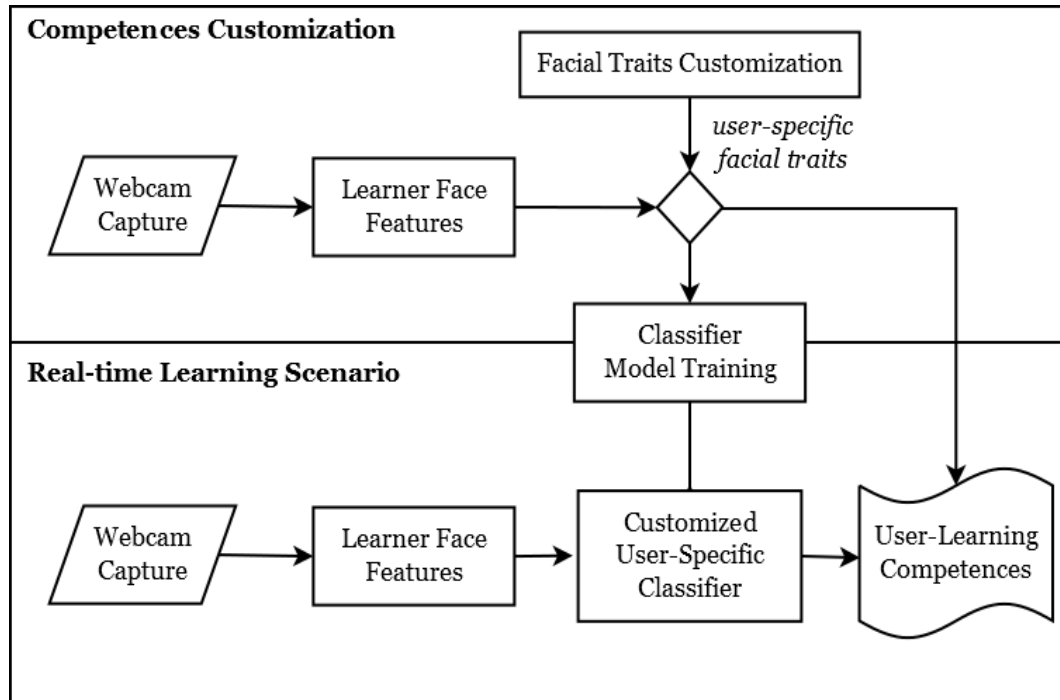
of classified facial emotions from different classroom activities. By statistical analysis (Torgo 2010) we will extract information such as the average emotion or stability of expressed emotions, sudden change of emotions, etc., which can be examined against simple parameters such as time, subject of study, and population. We expect to calibrate the background levels of facial emotion variance in our sample, and report the range in facial emotion trace between different learning scenarios. Potential yields would be a correlation method between a facial trait (eg boredom, engagement) and an output variable such as task performance. This could help to validate the hypothesis from video gaming that emotions can shape learning outcomes.

The 21<sup>st</sup> Century competencies are involved because our technology trial of the Geometrical Facial Emotions Classifier is taking place in a school which is actively and energetically implementing a 21<sup>st</sup> Century Skills curriculum at years 7-12. The school, in Porto, Portugal, offers a curriculum in two competencies (“grit” and “adaptability”) embedded in other disciplines. This has been piloted and will be in syllabus from September 2015 and therefore present in some of the classroom exercises where we are capturing facial emotion data. Assessment measures for these competencies have already been piloted at the school (self-assessment on a scale, and teacher/supervisor assessment, and peer assessment) and the progression data in these skills will be available at individual pupil level. Both datasets – facial emotions, and progressions in two 21<sup>st</sup> Century competencies – are potentially therefore available for comparison at cohort level. We will test data mining methods (e.g. variables covariance, correlation, scattering) [6] and our analysis may be able to identify paths for further empirical research or theorising around emotion patterns and 21<sup>st</sup> Century competency progression. Analytical methodologies for investigating non-cognitive attributes as predictors of outcomes has been established in fields such as population-scale psychometrics (Kosinski, Stillwell, and Graepel 2013) and to a lesser extent in education settings (Gray, Mc Guinness, and Owende, n.d.) and such approaches are assumed to be valid for testing correlations between facial emotion and output variables. With the small sample initially selected in our pilot school ( $n=50$  approx), the analysis is expected more to provoke methodological questions than provide robust insights, but the outcome should offer a preliminary baseline of quality and methodology for future data-driven research in emotion and competencies. If the area does look promising for further activity, one of its eventual fruits could be machine learning classifiers of facial emotion which detect learner emotion events and allow us to understand how these are linked to the attainment of learning competencies (Picard, Vyzas, and Healey 2001).

#### **4. Customisation of Facial Traits: User-specific Facial Emotion Tracking**

The diversity of faces, and the difference we assume to exist between learners’ competency progressions, will be a challenge to any generalised learning analytics involving facial emotion traces. Therefore, in addition to the cohort level study described above, our project will also research the question at the level of individual user facial traits in learning settings, using definition and customization and further tracking by CV algorithms (Szeliski 2010) as an extended application of our geometric algorithm (Loconsole et al. 2014). This user-

specific classifier has the important advantage it could examine the unique user in a wide range of learning contexts. It is not database dependent. Diagram 1 presents the methodology architecture.



**Diagram 1: Architecture of user-specific facial traits methodology.**

The method is divided in two consecutive steps. During step 1 (Customization), the learner face is captured by a webcam and user face features are tracked by a real-time face tracker [5]. Reference to Ekman values, and statistical analysis, can identify and classify significant moments in the learner facial emotion narrative. The teacher/supervisor also uses judgement to select further facial trait events which may be linked to important events learning narrative such as challenge, boredom, understanding. For example, the teacher/supervisor knows that learner Q blinks the left eye when he is nervous. Therefore, in this step the teacher/supervisor defines the blinking as facial trait. These two sources of insight create a custom set of user-specific facial traits. Our algorithm effectively uses the capture and facial trait defined in the Customization to train a user-specific machine learning classifier.

Step 2 is a real-time learning scenario in which a Customized User-Specific Classifier recognizes any incidence of a selected facial trait and stores this information in a database. The output would be that an individual learner has cascades of classifiers that are able to recognize his specific facial traits and link them with learning events. Since we are reducing the tracking to singular facial traits and to user-specific scenarios, the performance of the machine learning algorithm is maximised. Correlations can be examined at individual level for assessed performance in the 21<sup>st</sup> Century competencies (as measured at this school) and

also (for comparison) in performance assessments for traditional discipline-based skills and knowledge.

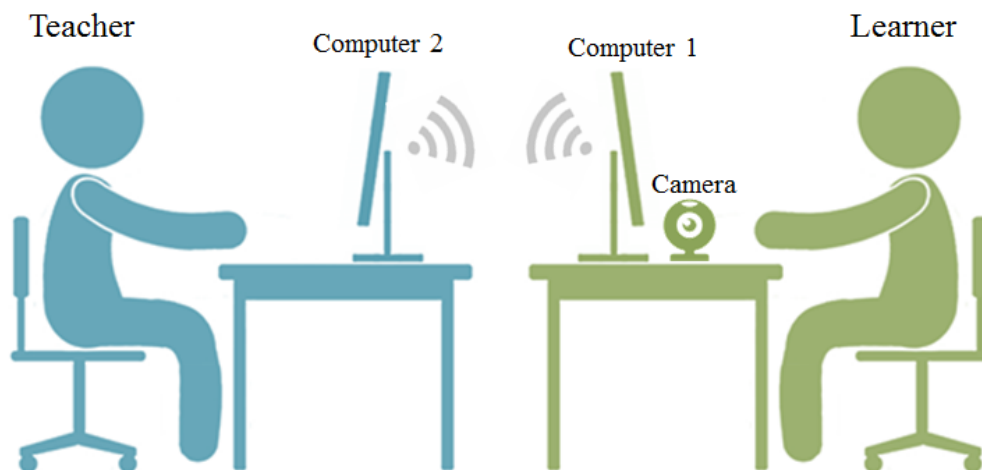
## 5. Technical Procedures for Acquiring Facial Emotion Data.

### *Facial Classifier Implementation*

In a screen-based learning exercise, the device webcam captures learner's facial movement data to send to the facial emotion classifier. Video is also captured. The supervisor/teacher screen allows them to direct the learner to different exercises according to the schedule of activities, and to make real time observations and annotations.

### *Customisation of Facial Traits*

Here we extend the previous set-up with a Customization step. In the same technical environment, the supervisor/teacher identifies a subset of facial expressions by direct observation of a learner's face directly from the video recording, alongside the facial expression data from the classifier in real time. This allows specific facial expressions for the learner to be tagged and recognised. When learning events with specific content (21<sup>st</sup> Century competency, or conventional syllabus) are taking place, the incidence of the custom facial expression can be tracked. Diagram 2 shows the learner setup is in green and the teacher set up in blue.



**Diagram 2: Technical procedure for customised user-specific facial expression tracking**

## 5. Discussion

The method we are reporting here should be viewed as a groundwork activity in establishing metrics that can inform 21<sup>st</sup> Century Skills learning analytics. If the method and the analysis suggest the facial emotions approach is feasible, the next step would be larger scale trial from 2016 onwards, exploring a range of cohorts and learning settings, tracking data over longer periods, and relating the data stream specifically to one or more of the commonly cited 21<sup>st</sup> Century skills.

While we are sanguine about the performance of facial emotion tracking as a technology in education settings, on the basis of its prior proof elsewhere, we note its known limitations in this application, which include:

- Facial occlusions (e.g. hand in front of the mouth) limit and influence negatively the accuracy of facial features tracking and emotion recognition
- Geometrical Recognition Algorithm using consumer level optical devices: Failure in detection of micro-expressions, since these have short durations and a standard optical device (i.e. webcam) cannot capture such facial movements
- We are only able to classify single emotions in the current experiment. Co-occurring different emotions in combined arrangements are not accessed.
- There is no scientific foundation, even at a theoretical level, on how to associate facial expressions to any learning process or competence (while the anecdotal frameworks implied by commonsense dicta such as “happy learners progress faster” may be valid, they are not proven by any robust standard.
- The patternings of facial movement/behaviour/pose may prove too complex or inconsistent to allow classifications that can be meaningfully mapped to any learning outcome.

We also highlight the weakness of the current understanding of what a learning analytics process for 21<sup>st</sup> Century Competencies would look like, even if challenges around capturing non-cognitive factors were solved.

Stanford’s Learning Analytics Workgroup (LAW) has described metrics of non-cognitive competencies as subject to processes of social evolution (Pea 2014) - there may be no stable framework for interpreting them as a data class. Meanwhile, education institutions are adopting non-cognitive skills metrics with a view to obtaining corporate advantage, creating de facto standards of measurement. For example, university admissions officers are reportedly selecting applicants on basis of non-cognitive skills evidence (“Noncognitive Measures: The Next Frontier in College Admissions” 2015)

The National Science Foundation’s Analytics for Learning project has highlighted that progress in analytics around non-cognitive competencies will require “attention to learning theory and merging multiple types of data” but has not, at time of this writing, provided analytics design patterns for the affective states it selected for non-cognitive skills analytics. (Bhanot 2015)

The Workshop *Non-Cognitive factors and Personalisation for Adaptive Learning* (NCFPAL) included one empirical assay of incorporating metrics of non-cognitive factors from the 21<sup>st</sup> Century skillset into a learning analytics framework (“Walkington\_etal\_EDM\_Readability\_Invited\_Paper,” n.d.). Gray’s study of Higher Education outcomes in a cohort of n=1207 (Gray, Mc Guinness, and Owende, n.d.) showed high correlations with GPA could be obtained for non-cognitive factors to do with approach and personality, when classification was performed using a k-Nearest Neighbour (k-NN) algorithm (suggesting non-linear models apply in the non-cognitive domain). However due

to overlapping constructs there was variation of significance between courses and a lack of consensus on the impact of her respective non-cognitive factors. Gray also demonstrated that the addition of non-cognitive constructs to cognitive variables did not improve predictive accuracy for GPA as a measure of educational success – suggesting that all the metrics were pointing to a similar underlying construct.

Against the background of the current weak or emerging understanding of the conceptual and theoretical basis of learning analytics for 21<sup>st</sup> Century skills, the value of a new biometrically based data class of facially expressed emotion is twofold. First, the nature of our capture process brings scalability and precision to the input data. Secondly, facial emotions access constructs that are potentially prior to and possibly not overlapping with culturally mediated constructs such as “growth mindset” or “problem framing”.

Ethical and acceptability issues around accessing information on learners’ emotions are explored along with the establishment of protocols for facial emotion recognition. The non-cognitive states implicit in the 21<sup>st</sup> Century Competency set are arguably distinctively interior, personal and prior domains, which are not normally assessed and tracked by institutions (as opposed to the more normally tracked outward manifestations of displayed ability at tasks). To that extent, we need to investigate learners’ views on revealing such private states in education processes. An acceptability survey as part of our work in the school (targeted at parents, teachers and learners) will offer initial data on this.

We look forward to critical feedback on our protocols and welcome applications from collaborators to test and reproduce or adapt the methodology in other settings, in the hope that new research lines in the field of behaviour tracking and analysis, based on facial emotion recognition.

## References

- Alves, Samanta, António Marques, Cristina Queirós, and Verónica Orvalho. 2013. “LIFEisGAME Prototype: A Serious Game about Emotions for Children with Autism Spectrum Disorders.” *PsychNology Journal* 11 (3): 191–211.
- Antonio, R. 1995. “Descartes’ Error: Emotion, Reason, and the Human Brain”. Avon Books.
- Bhanot, Ruchi. 2015. “A4L Design Patterns | A4L | Analytics for Learning.” Accessed July 10. <http://analytics4learning.org/a4l-design-patterns/>.
- Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, and National Research Council. 2011. *Assessing 21st Century Skills:: Summary of a Workshop*. National Academies Press.
- DiCerbo, Kristen E. 2014. “Game-Based Assessment of Persistence.” *Journal of Educational Technology & Society* 17 (1). International Forum of Educational Technology & Society: 17–28.
- Ekman, Paul, and Wallace V Friesen. 2003. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. ISHK.
- Famington, Camille A, Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W Johnson, and Nicole O Beechum. 2012. *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance: A Critical Literature Review*. Consortium On Chicago School Research.

- Gillies, Val. 2011. "Social and Emotional Pedagogies: Critiquing the New Orthodoxy of Emotion in Classroom Behaviour Management." *British Journal of Sociology of Education* 32 (2): 185–202.
- Gray, Geraldine, Colm Mc Guinness, and Philip Owende. n.d. "Non-Cognitive Factors of Learning as Predictors of Academic Performance in Tertiary Education." [http://ceur-ws.org/Vol-1183/ncfpal\\_paper06.pdf](http://ceur-ws.org/Vol-1183/ncfpal_paper06.pdf).
- Immordino-Yang, Mary Helen, and Joanna A Christodoulou. 2013. "Neuroscientific Contributions to Understanding and Measuring Emotions in Educational Contexts." In *International Handbook of Emotions in Education*. Routledge.
- Jamshidnezhad, A, and M Nordin. 2012. "Challenging of Facial Expressions Classification Systems: Survey, Critical Considerations and Direction of Future Work." *Research Journal of Applied Sciences* 4.
- Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): 5802–5.
- Loconsole, Claudio, Domenico Chiaradia, Vitoantonio Bevilacqua, and Antonio Frisoli. 2014. "Real-Time Emotion Recognition: An Improved Hybrid Approach for Classification Performance." In *Intelligent Computing Theory*, 320–31. Lecture Notes in Computer Science. Springer International Publishing.
- "'Noncognitive' Measures: The Next Frontier in College Admissions." 2015. *The Chronicle of Higher Education*. Accessed July 10. <http://chronicle.com/article/Colleges-Seek-Noncognitive/136621/>.
- OECD. 2013. "The Skills Needed for the 21st Century." In *OECD Skills Outlook 2013*, 45–54. OECD Publishing.
- "[PDF]New Vision for Education Report - Weforum.org - World Economic Forum." n.d. [http://www3.weforum.org/docs/WEFUSA\\_NewVisionforEducation\\_Report2015.pdf](http://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf).
- Pea, R. 2014. *A Report on Building the Field of Learning Analytics for Personalized Learning at Scale*. Stanford University.
- Pekrun, Reinhard, and Lisa Linnenbrink-Garcia. 2014. *International Handbook of Emotions in Education*. Routledge.
- Picard, R W, E Vyzas, and J Healey. 2001. "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State." *Pattern Analysis and Machine*. [ieeexplore.ieee.org](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=954607). [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=954607](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=954607).
- Shechtman, Nicole, Angela H DeBarger, Carolyn Dornsife, Soren Rosier, and L Yarnall. 2013. "Promoting Grit, Tenacity, and Perseverance: Critical Factors for Success in the 21st Century." *Washington, DC: US Department of Education, Department of Educational Technology*, 1–107.
- Szeliski, R. 2010. "Computer Vision: Algorithms and Applications". [books.google.com](https://books.google.com/books?hl=en&lr=&id=bXzAlkODwa8C&oi=fnd&pg=PR4&dq=Computer%2Bvision:%2Bgorithms%2Band%2Bapplications&ots=gZ_89ZmBBJ&sig=uJ-XhwWNTqmbOxeJg6dnlcU8rbA). [https://books.google.com/books?hl=en&lr=&id=bXzAlkODwa8C&oi=fnd&pg=PR4&dq=Computer%2Bvision:%2Bgorithms%2Band%2Bapplications&ots=gZ\\_89ZmBBJ&sig=uJ-XhwWNTqmbOxeJg6dnlcU8rbA](https://books.google.com/books?hl=en&lr=&id=bXzAlkODwa8C&oi=fnd&pg=PR4&dq=Computer%2Bvision:%2Bgorithms%2Band%2Bapplications&ots=gZ_89ZmBBJ&sig=uJ-XhwWNTqmbOxeJg6dnlcU8rbA).
- Torgo, Luis. 2010. *Data Mining with R: Learning with Case Studies*. Taylor & Francis.
- Ventura, Matthew, Valerie Shute, and Matthew Small. 2014. "--Assessing Persistence in Educational Games." *Design Recommendations for Intelligent Tutoring Systems*, 93.
- "Walkington\_etal\_EDM\_Readability\_Invited\_Paper." n.d. [http://ceur-ws.org/Vol-1183/ncfpal2014\\_proceedings.pdf](http://ceur-ws.org/Vol-1183/ncfpal2014_proceedings.pdf).



# References

- [AFB<sup>+</sup>13] Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, Mike Eheler, Zybnek Kysela, and Javier von der Pahlen. Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, SIGGRAPH '13, pages 1:1–1:1, New York, NY, USA, 2013. ACM.
- [AM06] Keith Anderson and Peter W McOwan. A real-time automated system for the recognition of human facial expressions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(1):96–105, 2006.
- [AMQO13] Samanta Alves, António Marques, Cristina Queirós, and Verónica Orvalho. Lifeisgame prototype: A serious game about emotions for children with autism spectrum disorders. *PsychNology Journal*, 11(3):191–211, 2013.
- [ARL<sup>+</sup>09] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: Photoreal facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, SIGGRAPH '09, pages 12:1–12:15, New York, NY, USA, 2009. ACM.
- [ASWG09] Akshay Asthana, Jason Saragih, Michael Wagner, and Roland Goecke. Evaluating AAM fitting methods for facial expression recognition. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009.
- [AW99] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

- [AZCP13] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3444–3451. IEEE, 2013.
- [Bag12] Daniel Lélis Baggio. *Mastering OpenCV with practical computer vision projects*. Packt Publishing Ltd, 2012.
- [BBRA<sup>+</sup>14] K Blom, A Bellido Rivas, Xenxo Alvarez, Ozan Cetinaslan, Bruno Oliveira, Verónica Orvalho, and Mel Slater. Achieving participant acceptance of their avatars. *Presence*, 23(3):287–299, 2014.
- [BCF04] Nuala Brady, Mark Campbell, and Mary Flaherty. My left brain and me: a dissociation in the perception of self and others. *Neuropsychologia*, 42(9):1156–1161, 2004.
- [BCF05] Nuala Brady, Mark Campbell, and Mary Flaherty. Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. *Brain and cognition*, 58(3):334–342, 2005.
- [BCL01] Fabrice Bourel, Claude C Chibelushi, and Adrian A Low. Recognition of facial expressions in the presence of occlusion. In *BMVC*, pages 1–10. Citeseer, 2001.
- [Bet09] V.K. Bettadapura. Face expression recognition and analysis: The state of the art. *Emotion*, pages 1–27, 2009.
- [Bet12] Vinay Bettadapura. Face expression recognition and analysis: the state of the art. *Course Paper, Visual Interfaces to Computer*, 2012.
- [BHB<sup>+</sup>11] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM, 2011.
- [BHPS10] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. *ACM Transactions on Graphics (TOG)*, 29(4):41, 2010.
- [Bio97] Frank Biocca. The cyborg’s dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, 3(2):0–0, 1997.

- [BKP05] Ioan Buciu, Irene Kotsia, and Ioannis Pitas. Facial expression analysis under partial occlusion. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–453. IEEE, 2005.
- [BLB<sup>+</sup>08] Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A Otaduy, and Markus Gross. Pose-space animation and transfer of facial details. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 57–66. Eurographics Association, 2008.
- [BLFM03] M.S. Bartlett, G. Littlewort, I. Fasel, and J.R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE, 2003.
- [BMT13] Sharon L Buxton, Lorraine MacDonald, and Lynette J Tippet. Impaired recognition of prosody and subtle emotional facial expressions in parkinson’s disease. *Behavioral neuroscience*, 127(2):193, 2013.
- [BMW<sup>+</sup>06] George Borshukov, Jefferson Montgomery, Witek Werner, Barry Ruff, James Lau, Paul Thuriot, Patrick Mooney, Stefan Van Niekerk, Dave Raposo, Jean-Luc Duprat, et al. Playable universal capture. In *ACM SIGGRAPH 2006 Sketches*, page 28. ACM, 2006.
- [BP14] Sofien Bouaziz and Mark Pauly. Semi-supervised facial animation retargeting. Technical report, 2014.
- [Bre01] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [Bré03] Serge Brédart. Recognising the usual orientation of one’s own face: the role of asymmetrically located details. *Perception*, 32(7):805–811, 2003.
- [BRM12] Tadas Baltrusaitis, Peter Robinson, and L Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2610–2617. IEEE, 2012.
- [BSSM<sup>+</sup>13] Dario Bombari, Petra C Schmid, Marianne Schmid Mast, Sandra Birri, Fred W Mast, and Janek S Lobmaier. Emotion recognition: The role of featural and configural face information. *The Quarterly Journal of Experimental Psychology*, 66(12):2426–2442, 2013.

- [BV99a] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.
- [BV99b] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.
- [CC] David Cristinacce and Tim Cootes. Feature Detection and Tracking with Constrained Local Models. *Biomedical Engineering*, pages 1–10.
- [CC92] T.F. Cootes and C.J.Taylor. Active shape models - smart snakes. In *In British Machine Vision Conference*, pages 266–275. Springer-Verlag, 1992.
- [CC08] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.
- [CDHRF08] Corrado Corradi-Dell’Acqua, Maïke D Hesse, Raffaella I Rumiati, and Gereon R Fink. Where is a nose with respect to a foot? the left posterior parietal cortex processes spatial relationships among body parts. *Cerebral cortex*, 18(12):2879–2890, 2008.
- [CET01] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.
- [CH67] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.
- [CHZ14] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43, 2014.
- [CI99] Jeffrey S Coffin and Darryl Ingram. Facial recognition system for security access and identification, November 23 1999. US Patent 5,991,429.
- [CK09] Yeongjae Cheon and Daijin Kim. Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7):1340 – 1350, 2009.

- [Cot10] Shane F. Cotter. Sparse Representation for accurate classification of corrupted and occluded facial expressions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 838–841. IEEE, 2010.
- [CS10] S. Chatterjee and H. Shi. A novel neuro fuzzy approach to human emotion determination. In *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, pages 282–287. IEEE, 2010.
- [CSG<sup>+</sup>03] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003.
- [CWLZ13] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41, 2013.
- [CWS<sup>+</sup>13] Yen-Lin Chen, Hsiang-Tao Wu, Fuhao Shi, Xin Tong, and Jinxiang Chai. Accurate and Robust 3D Facial Capture Using a Single RGBD Camera. In *2013 IEEE International Conference on Computer Vision*, pages 3615–3622. IEEE, December 2013.
- [CWWS12] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression, December 27 2012. US Patent App. 13/728,584.
- [CXH03] Jin-xiang Chai, Jing Xiao, and Jessica Hodgins. Vision-based control of 3d facial animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 193–206. Eurographics Association, 2003.
- [DB11] Christel Devue and Serge Brédart. The neural correlates of visual self-recognition. *Consciousness and cognition*, 20(1):40–51, 2011.
- [Deg88] B. Degraf. Mike the talking head. In *ACM Siggraph Electronic Theatre*, volume 3. ACM, 1988.
- [DHT<sup>+</sup>00] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field

- of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000.
- [DMB08] Ludovic Dutreve, Alexandre Meyer, and Saïda Bouakaz. Feature points based facial animation retargeting. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 197–200. ACM, 2008.
- [DMOB10] Ludovic Dutreve, Alexandre Meyer, Veronica Orvalho, and Saïda Bouakaz. Easy rigging of face by automatic registration and transfer of skinning parameters. In *Computer Vision and Graphics*, pages 333–341. Springer, 2010.
- [DN07] Zhigang Deng and Junyong Noh. Computer facial animation: A survey. In *Data-Driven 3D Facial Animation*, pages 1–28. Springer, 2007.
- [DWP10] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010.
- [EA11] Hedwig Eisenbarth and Georg W Alpers. Happy mouth and sad eyes: scanning emotional facial expressions. *Emotion*, 11(4):860, 2011.
- [EBDP96] Irfan Essa, Sumit Basu, Trevor Darrell, and Alex Pentland. Modeling, tracking and interactive animation of faces and heads//using input from video. In *Computer Animation’96. Proceedings*, pages 68–79. IEEE, 1996.
- [EF75] Paul Ekman and Wallace V Friesen. Unmasking the face: A guide to recognizing emotions from facial cues, 1975.
- [EF78] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.
- [Far00] Gunnar Farneback. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 135–139. IEEE, 2000.

- [Far03] Gunnar Farneback. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.
- [Fis04] Robert Fischer. Automatic Facial Expression Analysis and Emotional Classification by. *October*, 2004.
- [FL03] Beat Fasel and Juergen Luetttin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.
- [FMV<sup>+</sup>04] Rachel Falk, Denise Minter, Conrad Vernon, Guillaume Aretos, Lucia Modesto, Arnauld Lamorlette, Nick Walker, Tim Cheung, Janet Rentel-Lavin, and Harry Max. Art-directed technology: anatomy of a shrek2 sequence. In *ACM SIGGRAPH 2004 Course Notes*, page 13. ACM, 2004.
- [FPL<sup>+</sup>13] Christina T Fuentes, Mariella Pazzaglia, Matthew R Longo, Giorgio Scivoletto, and Patrick Haggard. Body image distortions following spinal cord injury. *Journal of Neurology, Neurosurgery & Psychiatry*, 84(2):201–207, 2013.
- [FRB<sup>+</sup>13] Christina T Fuentes, Catarina Runa, Xenxo Alvarez Blanco, Verónica Orvalho, and Patrick Haggard. Does my face fit?: A face image task reveals structure and distortions of facial feature representation. *PloS one*, 8(10):e76805, 2013.
- [Gel08] Tom Geller. Overcoming the uncanny valley. *Computer Graphics and Applications, IEEE*, 28(4):11–17, 2008.
- [GFPS11] Mar González-Franco, Tabitha C Peck, and Mel Slater. Virtual embodiment elicits a mu rhythm erd when the virtual hand is threatened. In *8th International Brain Research Organisation, Congress of Neuroscience*, 2011.
- [GOL] GOLEM. Golem: Realistic virtual humans.
- [GXhJlXg09] Lei Gang, Li Xiao-hua, Zhou Ji-liu, and Gong Xiao-gang. Geometric feature based facial expression recognition using multiclass support vector machines. In *Granular Computing, 2009, GRC '09. IEEE International Conference on*, pages 318 –321, aug. 2009.
- [Ham07] Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. *International Journal of Approximate Reasoning*, 46(3):542 – 567, 2007.

- [HCTW11] Haoda Huang, Jinxiang Chai, Xin Tong, and Hsiang-Tao Wu. Leveraging motion capture and 3d scanning for high-fidelity facial performance acquisition. *ACM Transactions on Graphics (TOG)*, 30(4):74, 2011.
- [HRBLM07a] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.
- [HRBLM07b] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [JA09] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *JIPS*, 5(2):41–68, 2009.
- [JJ13] Rachael E Jack and Rachael E Jack. Culture and facial expressions of emotion Culture and facial expressions of emotion. *Visual Cognition*, 00(00):1–39, 2013.
- [JN12] A. Jamshidnezhad and M.D.J. Nordin. Challenging of facial expressions classification systems: Survey, critical considerations and direction of future work. *Research Journal of Applied Sciences*, 4, 2012.
- [KB10] Helen Keyes and Nuala Brady. Self-face recognition is characterized by bilateral gain and by faster, more accurate performance which persists when faces are inverted. *The Quarterly Journal of Experimental Psychology*, 63(5):840–847, 2010.
- [KBP08] Irene Kotsia, Ioan Buciu, and Ioannis Pitas. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052 – 1067, 2008.
- [KGS12] Konstantina Kilteni, Raphaela Groten, and Mel Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012.
- [Kom88] Koji Komatsu. Human skin model capable of natural shape variation. *The visual computer*, 3(5):265–271, 1988.



- [KP07] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *Image Processing, IEEE Transactions on*, 16(1):172–187, jan. 2007.
- [KQP03] Ashish Kapoor, Yuan Qi, and Rosalind W. Picard. Fully automatic upper facial action recognition. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, AMFG '03, pages 195–, Washington, DC, USA, 2003. IEEE Computer Society.
- [KR12] ShashidharG. Koolagudi and K.Sreenivasa Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117, 2012.
- [KS10] K.E. Ko and K.B. Sim. Development of a facial emotion recognition method based on combining aam with dbn. In *Cyberworlds (CW), 2010 International Conference on*, pages 87–91. IEEE, 2010.
- [LA10] JP Lewis and Ken-ichi Anjyo. Direct manipulation blendshapes. *IEEE Computer Graphics and Applications*, 30(4):42–50, 2010.
- [LAR<sup>+</sup>14] JP Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. In *Eurographics 2014-State of the Art Reports*, pages 199–218. The Eurographics Association, 2014.
- [Las87] John Lasseter. Principles of traditional animation applied to 3d computer animation. In *ACM Siggraph Computer Graphics*, volume 21, pages 35–44. ACM, 1987.
- [LBFCO12] Claudio Loconsole, Nuno Barbosa, Antonio Frisoli, and Vernica Costa Orvalho. A new marker-less 3d kinect-based system for facial anthropometric measurements. In Francisco Perales, Robert Fisher, and Thomas Moeslund, editors, *Articulated Motion and Deformable Objects*, volume 7378 of *Lecture Notes in Computer Science*, pages 124–133. Springer Berlin / Heidelberg, 2012.
- [LBJ11] Y. Luximon, R. Ball, and L. Justice. The 3d chinese head and face modeling. *Computer-Aided Design*, 2011.
- [LCA05] Caroline Larboulette, Marie-Paule Cani, and Bruno Arnaldi. Dynamic skinning: adding real-time dynamic effects to an existing character

- animation. In *Proceedings of the 21st spring conference on Computer graphics*, pages 87–93. ACM, 2005.
- [LCF00] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000.
- [LCK<sup>+</sup>10] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.
- [LDBW] Oliver Langner, Ron Dotsch, Gijs Bijlstra, and Daniel Wigboldus. Support material for the article : Presentation and Validation of the Radboud Faces Database ( RaFD ) Mean Validation Data : Caucasian Adult Subset. *Image (Rochester, N.Y.)*.
- [Lew06] J P Lewis. Siggraph 2006 course notes Performance-driven Facial Animation Introduction. pages 1–5, 2006.
- [LH10] Matthew R Longo and Patrick Haggard. An implicit body representation underlying human position sense. *Proceedings of the National Academy of Sciences*, 107(26):11727–11732, 2010.
- [LH12] Matthew R Longo and Patrick Haggard. Implicit body representations and the conscious body image. *Acta psychologica*, 141(2):164–168, 2012.
- [LIF09] LIFEisGAME. Lifeisgame: Learning of facial emotions using serious games, 2009.
- [LM02] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. IEEE, 2002.
- [LTO<sup>+</sup>15] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance

- sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), July 2015.
- [LWHW12] Christian Lang, Sven Wachsmuth, Marc Hanheide, and Heiko Wersing. Facial communicative signals. *International Journal of Social Robotics*, 4(3):249–262, 2012.
- [LWP10] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *ACM Transactions on Graphics (TOG)*, 29(4):32, 2010.
- [LYYB13a] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4), July 2013.
- [LYYB13b] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42:1–42:10, July 2013.
- [Mar03] Chris Maraffi. *Maya character creation: modeling and animation controls*. New Riders, 2003.
- [Mar09] Clancy W Martin. *The philosophy of deception*. Oxford University Press, 2009.
- [McC93] Scott McCloud. Understanding comics: The invisible art. *Northampton, Mass*, 1993.
- [McC06] Scott McCloud. Making comics: Storytelling secrets of comics, manga and graphic novels author: Scott mccloud, publisher: William morrow. 2006.
- [MEK03] P. Michel and R. El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM, 2003.
- [Mor70] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.
- [MS07] Tim McLaughlin and Stuart S Sumida. The morphology of digital creatures. In *ACM SIGGRAPH*, pages 05–09, 2007.

- [MSK05] Daniel J Mundfrom, Dale G Shaw, and Tian Lu Ke. Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2):159–168, 2005.
- [MSSVT15] Lara Maister, Mel Slater, Maria V. Sanchez-Vives, and Manos Tsakiris. Changing bodies changes minds: owning another body affects social cognition. *Trends in Cognitive Sciences*, 19(1):6 – 12, 2015.
- [MTPT88] Nadia Magnenat-Thalmann, E Primeau, and Daniel Thalmann. Abstract muscle action procedures for human face animation. *The Visual Computer*, 3(5):290–297, 1988.
- [NAHF<sup>+</sup>12] R. Niese, A. Al-Hamadi, A. Farag, H. Neumann, and B. Michaelis. Facial expression recognition based on geometric and optical flow features in colour image sequences. *Computer Vision, IET*, 6(2):79–89, march 2012.
- [NN98] Jun-yong Noh and Ulrich Neumann. A survey of facial modeling and animation techniques. Technical report, USC Technical Report, 99–705, 1998.
- [OBP<sup>+</sup>12] Verónica Orvalho, Pedro Bastos, Frederic Parke, Bruno Oliveira, and Xenxo Alvarez. A facial rigging survey. In *33rd Annual Conference of the European Association for Computer Graphics-EUROGRAPHICS, May*, pages 13–18, 2012.
- [ope14] OpenCV, February 2014.
- [Orv07] Verónica Costa Teixeira Orvalho. *Reusable facial rigging and animation: Create once, use many*. PhD thesis, Universitat Politècnica de Catalunya, 2007.
- [Par72] Frederick I Parke. Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, pages 451–457. ACM, 1972.
- [PB02] M. Pardàs and A. Bonafonte. Facial animation parameters extraction and expression recognition using hidden markov models. *Signal Processing: Image Communication*, 17(9):675–688, 2002.
- [Per] PerceptionLab. [www.perceptionlab.com](http://www.perceptionlab.com).

- [PF03] Igor S Pandzic and Robert Forchheimer. *MPEG-4 facial animation: the standard, implementation and applications*. Wiley. com, 2003.
- [PMP<sup>+</sup>08] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.
- [PSAS13] Tabitha C Peck, Sofia Seinfeld, Salvatore M Aglioti, and Mel Slater. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and cognition*, 22(3):779–787, 2013.
- [PSS99] Frédéric Pighin, Richard Szeliski, and David H Salesin. Resynthesizing facial animation through 3d model-based tracking. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 143–150. IEEE, 1999.
- [PWA96] Frederic I Parke, Keith Waters, and Thomas R Alley. *Computer facial animation*, volume 55. AK Peters Wellesley, 1996.
- [RE01] Lionel Reveret and Irfan A Essa. Visual coding and tracking of speech related facial motion. 2001.
- [RKB12] Brendan Rooney, Helen Keyes, and Nuala Brady. Shared or separate mechanisms for self-face and other-face processing? evidence from adaptation. *Frontiers in psychology*, 3, 2012.
- [Rot82] Scott D Roth. Ray casting for modeling solids. *Computer graphics and image processing*, 18(2):109–144, 1982.
- [RPL10] J.D. Rodriguez, A. Perez, and J.A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575, 2010.
- [SAK04] H. Seyedarabi, A. Aghagolzadeh, and S. Khanmohammadi. Recognition of six basic facial expressions by feature-points tracking using rbf neural network and fuzzy inference system. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 2, pages 1219 –1222 Vol.2, june 2004.
- [Sco93] McCloud Scott. Understanding comics: the invisible art, 1993.

- [SGBP04] Saurabh Singh, Aglika Gyaourova, George Bebis, and Ioannis Pavlidis. Infrared and visible image fusion for face recognition. In *Defense and Security*, pages 585–596. International Society for Optics and Photonics, 2004.
- [SGM09] C. Shan, S. Gong, and P.W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.
- [Sla14] Mel Slater. Grand challenges in virtual environments. *Frontiers in Robotics and AI*, 1:3, 2014.
- [SLC11a] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.
- [SLC11b] J.M. Saragih, S. Lucey, and J.F. Cohn. Real-time avatar animation from a single image. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 117–124. IEEE, 2011.
- [SSMCP02] J Schleifer, R SCADUTO-MENDOLA, Y CANETTI, and M PIRETTI. Character setup from rig mechanics to skin deformations: A practical approach. In *Proc. SIGGRAPH*, volume 2, 2002.
- [SSV14] Mel Slater and Maria V Sanchez-Vives. Transcending the self in immersive virtual reality. *Computer*, 47(7):24–30, 2014.
- [SZG15] Shuli Sun, Minglei Zhang, and Zhihong Gou. Smoothing algorithm for planar and surface mesh based on element geometric deformation. *Mathematical Problems in Engineering*, 2015, 2015.
- [Tia04] Ying-li Tian. Evaluation of face resolution for expression analysis. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW’04. Conference on*, pages 82–82. IEEE, 2004.
- [UKMS<sup>+</sup>05] Lucina Q Uddin, Jonas T Kaplan, Istvan Molnar-Szakacs, Eran Zaidel, and Marco Iacoboni. Self-face recognition activates a frontoparietal mirror network in the right hemisphere: an event-related fmri study. *Neuroimage*, 25(3):926–935, 2005.

- [VBPP05] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433. ACM, 2005.
- [vdPJD<sup>+</sup>14] Javier von der Pahlen, Jorge Jimenez, Etienne Danvoye, Paul Debevec, Graham Fyffe, and Oleg Alexander. Digital ira and beyond: creating real-time photoreal digital actors. In *ACM SIGGRAPH 2014 Courses*, page 1. ACM, 2014.
- [VER10] VERE. Vere: Virtual embodiment and robotic re-emobidment, 2010.
- [War04] Antony Ward. *Game character development with maya*. New Riders, 2004.
- [WBLP11] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. *ACM Transactions on Graphics (TOG)*, 30(4):77, 2011.
- [WC03] Tamara L Watson and Colin WG Clifford. Pulling faces: An investigation of the face-distortion aftereffect. *Perception-London*, 32(9):1109–1116, 2003.
- [Wil90] Lance Williams. Electronic mask technology. In *ACM SIGGRAPH'90 course notes*, page 10. ACM, 1990.
- [WLVGP09] Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '09, pages 7–16, New York, NY, USA, 2009. ACM.
- [WSE03] Lawrence B Wolff, Diego A Socolinsky, and Christopher K Eveland. Quantitative measurement of illumination invariance for face recognition using thermal infrared imagery. In *International Symposium on Optical Science and Technology*, pages 140–151. International Society for Optics and Photonics, 2003.
- [WY07] J. Wang and L. Yin. Static topographic modeling for facial expression recognition and analysis. *Computer Vision and Image Understanding*, 108(1-2):19–34, 2007.
- [XDIT13] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern*

- Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.
- [YA11] Aliaa A. A. Youssif and Wesam A. A. Asker. Automatic facial expression recognition system based on geometric and appearance features. *Computer and Information Science*, pages 115–124, 2011.
- [YZ06] Xiaosong Yang and Jian J Zhang. Stretch it-realistic smooth skinning. In *Computer Graphics, Imaging and Visualisation, 2006 International Conference on*, pages 323–328. IEEE, 2006.
- [ZH04] Song Zhang and Peisen Huang. High-resolution, real-time 3d shape acquisition. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 28–28. IEEE, 2004.
- [ZPRH09] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.
- [ZTC12] L. Zhang, D. Tjondronegoro, and V. Chandran. Discovering the best feature extraction and selection algorithms for spontaneous facial expression recognition. *2012 IEEE International Conference on Multi-media and Expo*, 2012.