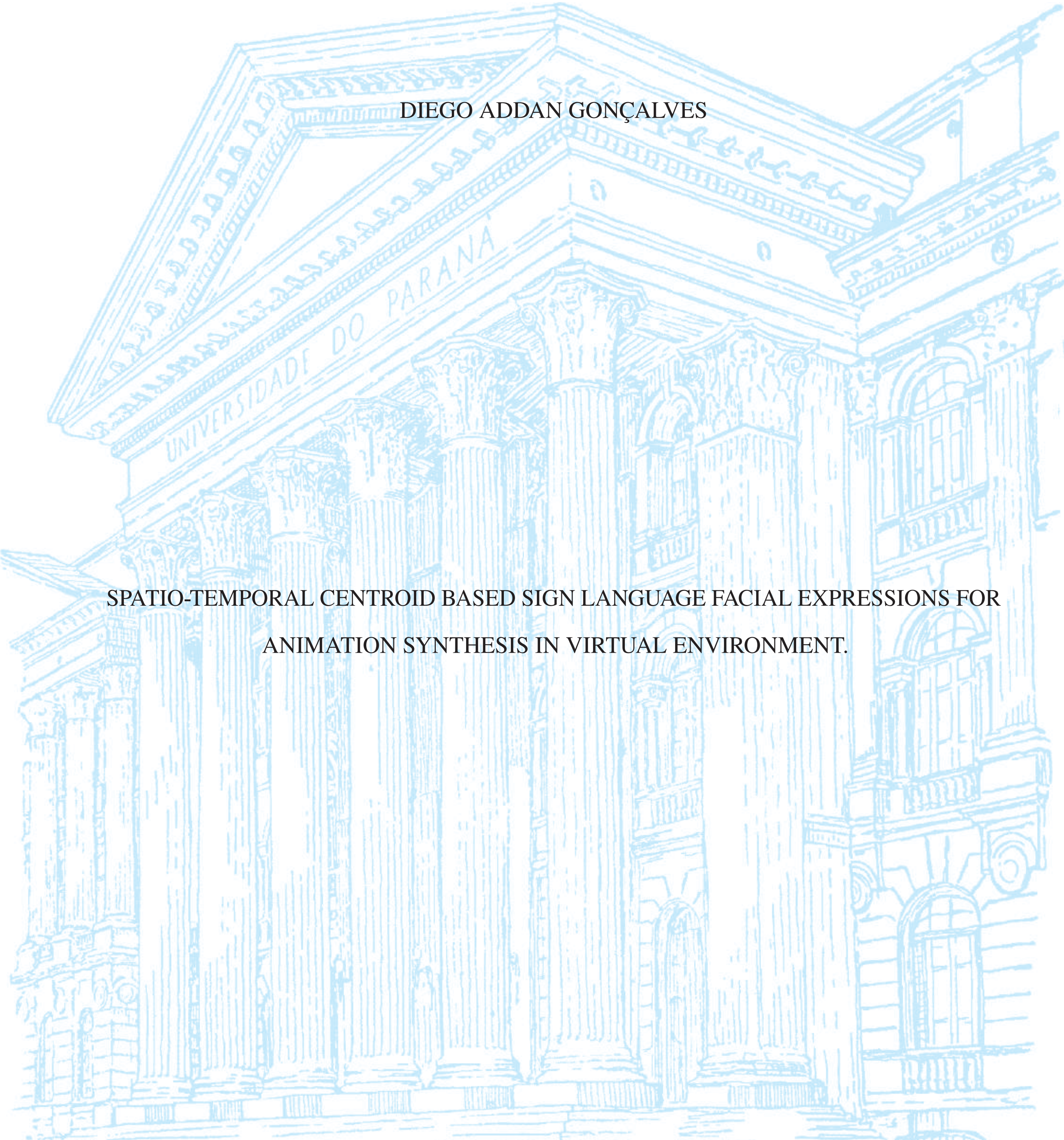
UNIVERSIDADE FEDERAL DO PARANÁ



DIEGO ADDAN GONÇALVES

SPATIO-TEMPORAL CENTROID BASED SIGN LANGUAGE FACIAL EXPRESSIONS FOR ANIMATION SYNTHESIS IN VIRTUAL ENVIRONMENT.

CURITIBA PR

2019

DIEGO ADDAN GONÇALVES

SPATIO-TEMPORAL CENTROID BASED SIGN LANGUAGE FACIAL EXPRESSIONS FOR ANIMATION SYNTHESIS IN VIRTUAL ENVIRONMENT.

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Ciência da Computação no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Eduardo Todt.

CURITIBA PR

2019

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

G635s Gonçalves, Diego Addan
Spatio-temporal centroid based sign language facial expressions for animation synthesis in virtual environment [Recurso eletrônico] / Diego Addan Gonçalves – Curitiba, 2019.

Tese - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Informática.
Orientador: Eduardo Todt.

1. Avatar 3 D. 2. Língua Brasileira de Sinais (LIBRAS). 3. Expressões faciais. I. Universidade Federal do Paraná. II. Todt, Eduardo. III. Título.

CDD: 006.693

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

MINISTÉRIO DA EDUCAÇÃO
SETOR SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA -
40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **DIEGO ADDAN GONÇALVES** intitulada: **Spatio-Temporal Centroid Based Sign Language Facial Expressions for Animation Synthesis in Virtual Environment**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 20 de Fevereiro de 2019.

EDUARDO TODT
Presidente da Banca Examinadora

Marc S
MARC STAMMINGER
Avaliador Externo (FAU)

Felipe
TANYA AMARA FELIPE DE SOUZA
Avaliador Externo (INES)

Laura S
LAURA SANCHEZ GARCIA
Avaliador Interno (UFPR)

BRUNO MÜLLER JUNIOR
Avaliador Externo (UFPR)

ACKNOWLEDGEMENTS

H. P. Lovecraft presented me with almost physical (and ironically impossible) description of Space-Time concepts that were then fixed in my life, reinforced by Robert Heinlein in his tale in Time Probe, and philosophically by Alan Moore in his book Jerusalem. Understanding and using those concepts have become an attractive challenge that remains until today, and that motivated the beginning of this research. Almost five years later, I see myself at the end of this journey, and the beginning of a new one, with a mixed sense of accomplishment and that the work must continue.

It was an honor to have Professor Eduardo Todt as advisor. The best teacher in the world, who taught me important professional and personal values. Thank you to all who took part in this journey. To the colleagues of the group "Vision, Robotics and Image" of UFPR for the conversations and friendship, to the students and professors of the group "Lehrstuhl Graphische Datenverarbeitung" of Friedrich-Alexander-Universität for welcoming me. Thanks to the teachers of the thesis tribunal, Marc Stamminger, Bruno Muller, Laura Sanchez and Tanya Felipe for the contributions.

To my love Aldiny Paula who was by my side and accompanied everything with patience and love, always putting me up. Thank you also to her mother Andreia and the whole family for their continued support. To my friend Michel, for reminding me of the way and of what matters. My mother Sandra and my brother Douglas, my best friend always. Love you all. Finally, I dedicate this thesis to my father José Gonçalves Neto, who taught me and inspired me eternally, and continues with me in some way ...

"In the highest level a man has the look of knowing nothing ."
Tsunetomo Yamamoto, Hagakure: The Book of the Samurai.

RESUMO

Formalmente reconhecida como segunda língua oficial brasileira, a BSL, ou Libras, conta hoje com muitas aplicações computacionais que integram a comunidade surda nas atividades cotidianas, oferecendo intérpretes virtuais representados por avatares 3D construídos utilizando modelos formais que parametrizam as características específicas das línguas de sinais. Estas aplicações, contudo, ainda consideram expressões faciais como recurso de segundo plano em uma língua primariamente gestual, ignorando a importância que expressões faciais e emoções imprimem no contexto da mensagem transmitida. Neste trabalho, a fim de definir um modelo facial parametrizado para uso em línguas de sinais, um sistema de síntese de expressões faciais através de um avatar 3D é proposto e um protótipo implementado. Neste sentido, um modelo de landmarks faciais separado por regiões é definido assim como uma modelagem de expressões base utilizando as bases faciais AKDEF e JAFEE como referência. Com este sistema é possível representar expressões complexas utilizando interpolação dos valores de intensidade na animação geométrica, de forma simplificada utilizando controle por centroides e deslocamento de regiões independentes no modelo 3D. É proposto ainda uma aplicação de modelo espaço-temporal para os landmarks faciais, com o objetivo de observar o comportamento e relação dos centroides na síntese das expressões base definindo quais pontos geométricos são relevantes no processo de interpolação e animação das expressões. Um sistema de exportação dos dados faciais seguindo o formato hierárquico utilizado na maioria dos avatares 3D intérpretes de línguas de sinais é desenvolvido, incentivando a integração em modelos formais computacionais já existentes na literatura, permitindo ainda a adaptação e alteração de valores e intensidades na representação das emoções. Assim, os modelos e conceitos apresentados propõe a integração de um modeo facial para representação de expressões na síntese de sinais oferecendo uma proposta simplificada e otimizada para aplicação dos recursos em avatares 3D.

Palavras-chave: Avatar 3D, Dados Espaço-Temporal, Libras, Língua de sinais, Expressões Faciais

ABSTRACT

Formally recognized as the second official Brazilian language, BSL, or Libras, today has many computational applications that integrate the deaf community into daily activities, offering virtual interpreters represented by 3D avatars built using formal models that parameterize the specific characteristics of sign languages. These applications, however, still consider facial expressions as a background feature in a primarily gestural language, ignoring the importance that facial expressions and emotions imprint on the context of the transmitted message. In this work, in order to define a parametrized facial model for use in sign languages, a system of synthesis of facial expressions through a 3D avatar is proposed and a prototype implemented. In this way, a model of facial landmarks separated by regions is defined as a modeling of base expressions using the AKDEF and JAFEE facial bases as a reference. With this system it is possible to represent complex expressions using interpolation of the intensity values in the geometric animation, in a simplified way using control by centroids and displacement of independent regions in the 3D model. A spatial-temporal model is proposed for the facial landmarks, with the objective of define the behavior and relation of the centroids in the synthesis of the basic expressions, pointing out which geometric landmark are relevant in the process of interpolation and animation of the expressions. A system for exporting facial data following the hierarchical format used in most avatars 3D sign language interpreters is developed, encouraging the integration in formal computer models already existent in the literature, also allowing the adaptation and change of values and intensities in the representation of the emotions. Thus, the models and concepts presented propose the integration of a facial model to represent expressions in the synthesis of signals offering a simplified and optimized proposal for the application of the resources in 3D avatars.

Keywords: 3D Avatar, Spatio-Temporal Data, BSL, Sign Language, Facial Expression

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

BSL	Brasilian Sign language
MPEG-4	Facial Animation Standard
XML	Extensible Markup Language
STIP	Spatio Temporal Interest Point
RGB-D	Color pattern with Depth channel
AKDEF	Averaged Karolinska Directed Emotional Faces
JAFEE	Japanese Female Facial Expression
PCA	Principal Components Algorithm
FA	Factor Analysis Algorithm
3DED	Three-dimensional Euclidean Distance
EDVA	Euclidean Distance Variance Analysis Algorithm
EDMA	Euclidean Distance Matrix Analysis
FPS	Frames Per Second
FEL	Facial Expression Landmark
UFPR	Universidade Federal do Paraná
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CONTENTS

1 INTRODUCTION

Systems that use virtual environments for information transmission are of fundamental importance in everyday life (Kacorri, 2015) (Tanaka et al., 2016). A constant concern in this area is to integrate persons with special needs allowing them to share information with everyone. For the deaf community, which has its own sign languages, virtual broadcasters representing messages by symbols and gestures contribute to this group integration into society, providing flexibility and speed in the transmission of information (Lombardo et al., 2011) (Punchimudiyanse e Meegama, 2015).

Although sign languages have been documented since the 17th century, their practical definitions and modeling vary locally based on countries legislation (Lombardo et al., 2011) (Wantroba e Romero, 2015) (Kacorri et al., 2015) (Bento et al., 2014), most of which based on recent studies. Brazilian Sign Language (BSL), for example, was formalized in 2002 when research's involving its parameters and formal definitions gained a positive impulse (Sofiato, 2014).

There are several Sign Languages computational synthesis systems developed around the world, based on signal synthesis through user interaction (Bento et al., 2014) (Grif e Manueva, 2016) (Kacorri, 2013) (Adhan e Pintavirooj, 2016) (Ratan e Hasler, 2014), (Kacorri et al., 2015). In general, these synthesis systems use configuration parameters for hand gestures, besides the body and arms positioning, aiming fidelity between the virtual and the real representations (Ferstl et al., 2016) (Rieger, 2003). It is possible to observe the predominant lack of facial expressions, although these features provide message context and define part of the information transmitted (Elons et al., 2014).

The accurate transmission of a message using sign languages really needs facial expression as a feeling modifier or as context supplement for the raw gesture information. In this context, facial expressions are representations of emotions such as joy or anger, or as helpers, such as questioning expression or irony. According to Neidle et al. (1998), it is remarked that an addressee in a sign language dialogue tends to look more to the eyes of the partner than to the hands, reinforcing the importance of facial expressions in the communication. The main challenge in the representation of facial expressions is the animation of a great number of details, since a facial deformation, even a small one, can change the message semantics, for example, from a positive expression of surprise to a negative of fear.

This becomes even more complex when considering the context of virtual environments where all details should be built aiming at realism and naturalness. In the process of signals synthesis for sign language, facial details follow models that define features and parameters (Kacorri e Huenerfauth, 2014) (Kacorri et al., 2015) indicating the position of controllers associated to joint elements, in addition to modeling the representation of non-manual parameters.

The rendered animations usually have a heavier computational cost related to the face and hands because it is where the details are concentrated (Wiegand, 2014). This problem is even greater considering that the face has more independent points that the hand and failure to get a realistic animation is more perceptible. In addition to displacements in the geometric mesh features such as textures should be considered beyond the focus on fine detail. Besides enhancing communication, facial expressions are considered very important in order to a robot or avatar get accepted by humans (Wantroba e Romero, 2015).

This work proposes and validates a novel parametric model of facial expression applied to a 3D avatar dedicated to sign language synthesis. A model that parameterizes the face mapping features that allow the representation of expressions in a controlled way is relevant for the complete transmission of the sign language message. The synthesis of facial expressions can be controlled by understanding the behavior of facial regions, defined by landmarks, along with geometric animations facilitating the implementation of automatic signal synthesis. 3D Avatar face is sectored, allowing a finer and more precise control of synthesis contributing to the implementation of more realistic and relevant animations for a certain expression and message. Throughout this research, models and methods for spatio-temporal data representation were proposed and applied, defining a model for the synthesis of basic facial expressions, also allowing the interpolation of derived expressions. The specific objectives and methodology are presented in the following sections.

1.1 CHALLENGES AND MOTIVATION

A recurring complicator on systems which use virtual environments is the computational cost, usually high because of 3D renders requirements. Elements that make these complex calculations are tied to a number of polygons involved, collisions, textures and effects such as particles or physical simulation (Azahar et al., 2010) (Streuber et al., 2016). As a consequence, there is a strong demand for new procedures to optimize the animation and output process. Regarding simulation of facial expressions in sign language, animation parameters should be considered and parametrized, opening another research front.

It is fundamental that there is a facial model with well defined computational parameters, besides a study that explains the behavior of these parameters, where the application of this model is simplified, encouraging its use in any model that uses spatial hierarchical parameters in order to define signals and their representations. BSL does not have a computational model specifically defined for the 3D avatar face that follows sectorized parameters, which would aid the fine control of the geometric deformations in the representation of emotions during the transmission of the message, defining which parameters are most relevant in the representation of the main expressions, allowing local control for each parameter.

Studies on gesture synthesis on 3D space have the immediate impact on the development of new essential technologies and deserve attention and investment (Han, 2015). In this direction, systems that use virtual interpreters need a greater focus on the parametrized representation of facial expressions, which is an indispensable element in the effective transmission of a message (Hyde et al., 2016). Thus, the main issues in the process of synthesis using avatars is the use of well-defined parameters in addition to fine detail control in the 3D mesh.

Facial expressions compose a significant part in non-verbal communication and they are essential for effective Human-Computer Interaction (Xue et al., 2014). According to (Neidle et al., 1998), a human face is the region with more fine and smaller details. During a conversation in sign language, the focus of attention is fixed on the partner's face, relegating hand gestures to peripheral vision or secondary attention. Otherwise, it could suffer a loss of relevant information once the face contains less noticeable but important details compared to variation in volume and timing in spoken languages (Wiegand, 2014).

The main approach of this work is to propose a functional facial model for Brazilian Sign Language (BSL) 3D animation that considers specific parameters of the face using spatio-temporal information. BSL has no formal descriptions for a facial model and there are few studies in the literature where facial landmarks are pointed and associated with basic emotions. This work aims to model the behavior of facial landmarks and to propose a model where facial expressions can be generated in an optimized way in order to encourage systems that already use 3D avatars for the transmission of signals in sign languages to integrate more complex facial expressions.

1.2 OBJECTIVES AND CONTRIBUTION

The objective of this work it is to define and apply a landmark-based parametrization model of facial expressions for use in a signal language synthesis system, in order to enhance the animation process of current systems, which will be integrated in the HCI-SL architecture module (Figure 1.1) that proposes the complete synthesis of BSL through 3D avatars CORE-SL that have mapped all manual parameters except the face (Iatskiu et al., 2017) (Gonçalves et al., 2015), supporting a more accurate representation of sign language messages through a 3D avatar. The main facial points relevant to the expressions representation on a 3D mesh are identified together with their spatial trajectory when performing animations associated with emotions.

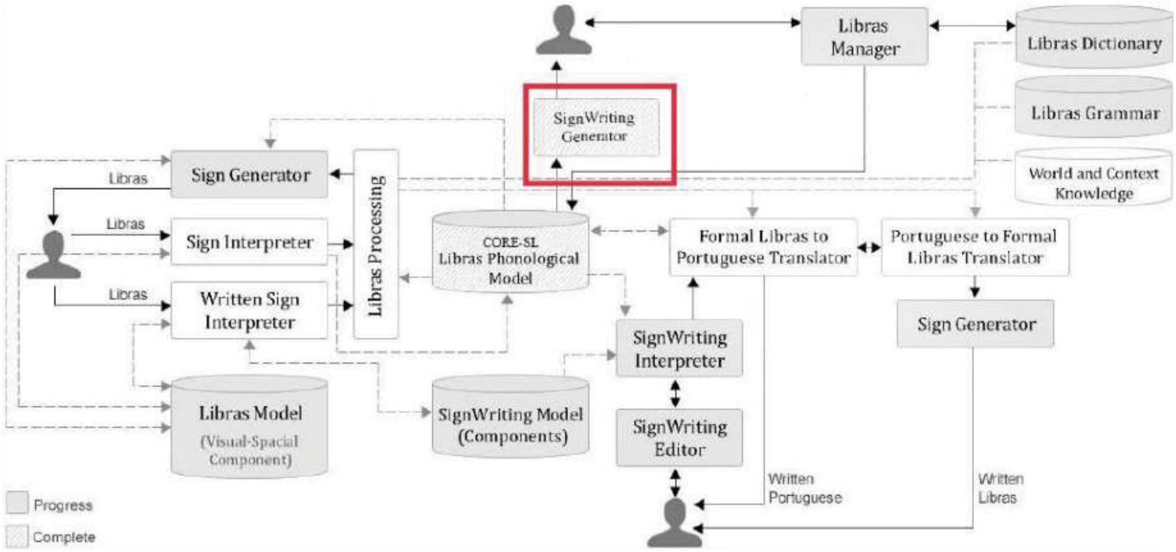


Figure 1.1: HCI-SL architecture module for BSL processing automatic translation and synthesis of the signals (Iatskiu et al., 2017). The facial parameters and synthesis can be integrated in the output process.

A formal model is proposed with the classification of emotions or expressions characteristics in BSL as well an optimized process to apply landmarks displacement from regions of the face in the synthesis of facial emotions.

The specific objective follows:

- Propose a parameterized model for facial representation in BSL.
- The identification of the principal components in the facial regions landmarks deformation, in order to select the most affected areas in an avatar during the synthesis process in a virtual environment. This makes possible to generate the interpolation of synthesized emotions in independent areas, so that the less important components for the synthesis of facial expression may be discarded, reducing the total computational cost necessary for the animation.
- Optimization of the animation generation based on the features and parameters involved in the facial animation of a 3D avatar.
- Exploration of temporal data related to a geometric mesh in order to control and optimize 4D actions of a 3D avatar.
- Validation of the proposed system through a dataset based on interpolation of parameters and generation of simplified and adapted expressions.

In Figure1.2 an overview of the complete process is shown. The Facial Parameters block contains a representation of the facial landmarks on the 3D mesh, obtained from the MPEG4-FP model and the definition of the regions of interest, as well a representation for emotions and expressions. The Landmarks Behavior block generates the animation, exporting the commands as an XML external file mapped as sign language formal models, based on the neutral expressions over the rigged face. The animation produces movements corresponding to the landmarks selected by a behavior analysis previously performed on the facial animations modeled. From this step, it is possible to optimize the process of interpolation of expressions, through the analysis of the behavior of the landmarks during the movements associated with the expressions.

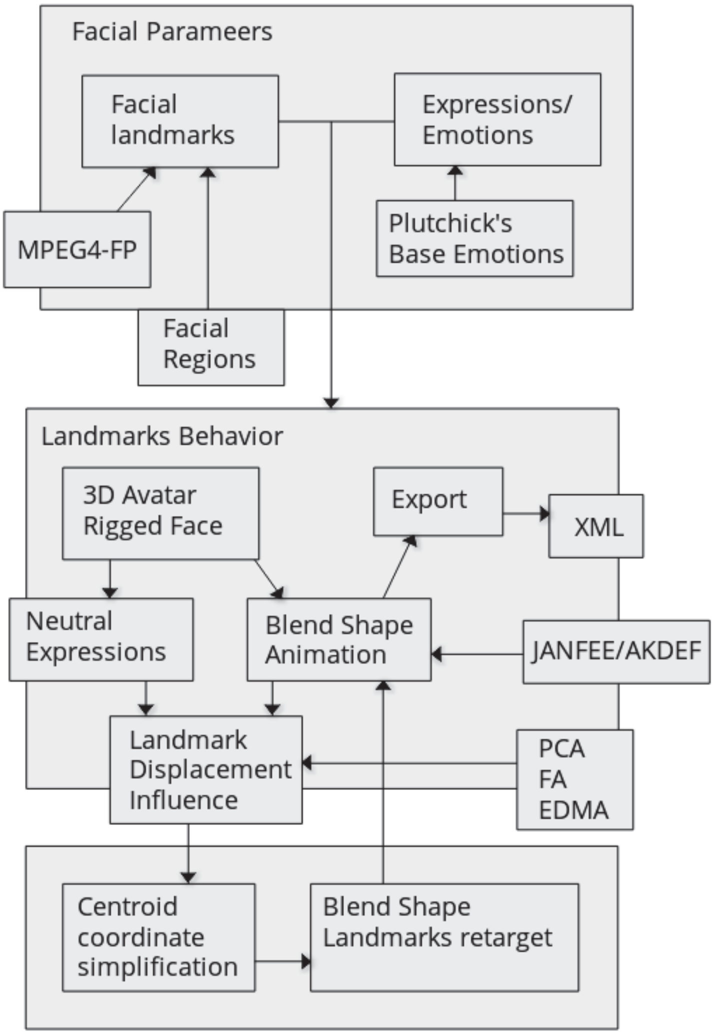


Figure 1.2: Overview of the animation process. There are two blocks: the Facial Parameters and the Landmarks Behavior. The first one has the facial landmarks associated to regions of interest to be animated as well as a representation for emotions and expressions to be reproduced. The Landmarks behavior is responsible for the animation generation. The process of simplifying and re-targeting of the coordinates and values of the facial landmarks are in an independent block because they can be executed externally by changing the textual file referring to the formal model.

The contributions are the definition of a process to built and synthesize emotions and expressions in a 3D avatar for BSL systems besides the extraction and use of spatio-temporal data of facial animation. It's defined as a facial model that considers simplified landmarks and regions for the representation of human base emotions. This formalizes an important element of the Signal Languages defining a computational facial model for BSL.

1.3 METHODOLOGY

The first step in this research is the investigation on the techniques implemented in virtual environments for signal synthesis, formal models for avatar facial representation defining landmarks in a geometric mesh, parametrized ways to categorize and represent these points and regions. Next, a non-manual model for BSL was developed. The model was constructed aiming the integration into generic sign language synthesis systems, including data parameters regarding face, and a parser method.

A 4D data model was defined aiming to classify the extracted information of the geometric mesh and subsequently optimize the animation process through landmarks displacement simplification in geometric surfaces. Thus, with the spatio-temporal model applied to the facial points of interest, it was possible to define the geometric behavior in the synthesis of the base expressions. Through this process, the facial model will be applied based on the formal computational models used to define the sign language hand parameters, applying its rules in the generation of facial animation.

1.4 OUTLINE

The next chapter presents related works, divided into three main areas: Virtual agents as an interpreter in systems for sign languages representation, facial expressions, and 4D data. These themes are related in the stage of practical experiments where the 4D modeling will be applied to the facial model for the synthesis of BSL through a 3D avatar. In Chapter 3 the facial model for use in Sign Language is defined followed by experiments in order to validate the Landmarks and proposed parameters.

Chapter 4 presents the system implemented for the extraction of facial data for an external model approaching the format used in the representation of hand parameters in signal languages. Next, in Chapter 5, the definitions of the 4D data used for the extraction and behavioral analysis by trajectories of the synthesis of the facial expressions in the Sign Language environment are presented. Following, in Chapter 5 is presented the validation and a general discussion of the experiments and results.

Final considerations, and perspectives for future work, besides the used bibliography conclude this research.

2 PREVIOUS WORKS

This chapter aims to explore related research in the areas approached by this work, which are: 3D avatars for sign languages synthesis, facial parameters, spatio-temporal data, and animation optimization methods.

These subjects repeatedly intersect in works where the scope involves animation in virtual environments (Alkawaz e Basori, 2012) (Kocon, 2014) (Kacorri, 2015) (Polys et al., 2018) (Basawapatna et al., 2018) (Bouzid et al., 2013) . The first sub-section focuses on works concerning the use of 3D avatars oriented to the representation of sign languages in virtual environments features used to represent facial expressions and computational representation of emotions.

2.1 VIRTUAL INTERPRETERS FOR SIGN LANGUAGES.

The use of virtual agents for educational or entertainment systems has increased as well as the interest by target users (Grif e Manueva, 2016) (Ratan e Hasler, 2014) (Wiegand, 2014) (Polys et al., 2018). Basawapatna et al. (2018) reinforce the importance of 3D avatars in educational environments, comparing the impact of more dynamic virtual environments with traditional concepts in programming teaching. In this direction, systems that use virtual interpreters need a greater focus on the parameterized representation of facial expressions, which is an indispensable element in the effective transmission of a message (Hyde et al., 2016). 3D avatars have an influence on learning, facilitating empathy between those who are watching and the announcer. Also, humanoid avatars are more relevant in instructional contexts and soft features on the avatar face would increase the positive impact on a large group of people (Draman e Zeki, 2014) Hada et al. (2018). Wauck et al. (2018) state that the effect of humanoid avatars is important for the user's empathy, although the physical similarity between user and avatar does not influence the experience itself. One of the ways to synthesize a sign language animation is through graphics engines which provide physical simulators used in a sign languages dynamic recognition and representation (Tangsuksant et al., 2014).

Kacorri et al. (2015) applied a questionnaire to users with hearing impairment in order to evaluate the acceptance of virtual interpreters for the representation of sign language gestures and their most important features. The importance of these systems has been observed as well their main weakness was pointed as being the lack of fidelity between the virtual and the real representations, especially on small details on animations. This is in part due to the lack or weak representation of facial expressions, since these features provide relevant message context and intensity modifiers (Elons et al., 2014) (Ratan e Hasler, 2014). Neidle et al. (1998) emphasize that in a sign language dialogue the addressee tends to look more to the eyes of the partner than to the hands, reinforcing the importance of facial expressions in the communication.

Computer animation is a sequential rendering of objects usually controlled by physical and mathematical simulation. Current approaches for obtaining more realistic animations of the avatars use motion capture or controllers manipulation with tracking of landmarks in videos of real interpreters or extracting depth coordinates using capture sensors (Bento et al., 2014) (Ahire et al., 2015) as shown in Figure 2.1. The main challenge of a more realistic virtual animation relies in the details. A more natural synthesis process demands greater computational cost, considering specific parameters and controllers, complex geometry mesh and massive calculation. There are examples of avatars in two dimensions, which require less computational cost but offer a more limited level of animations or realism Sahid et al. (2016). There is no sign language avatar that has animations with fine control of facial expression or animations made with realism techniques such as rotoscoping.

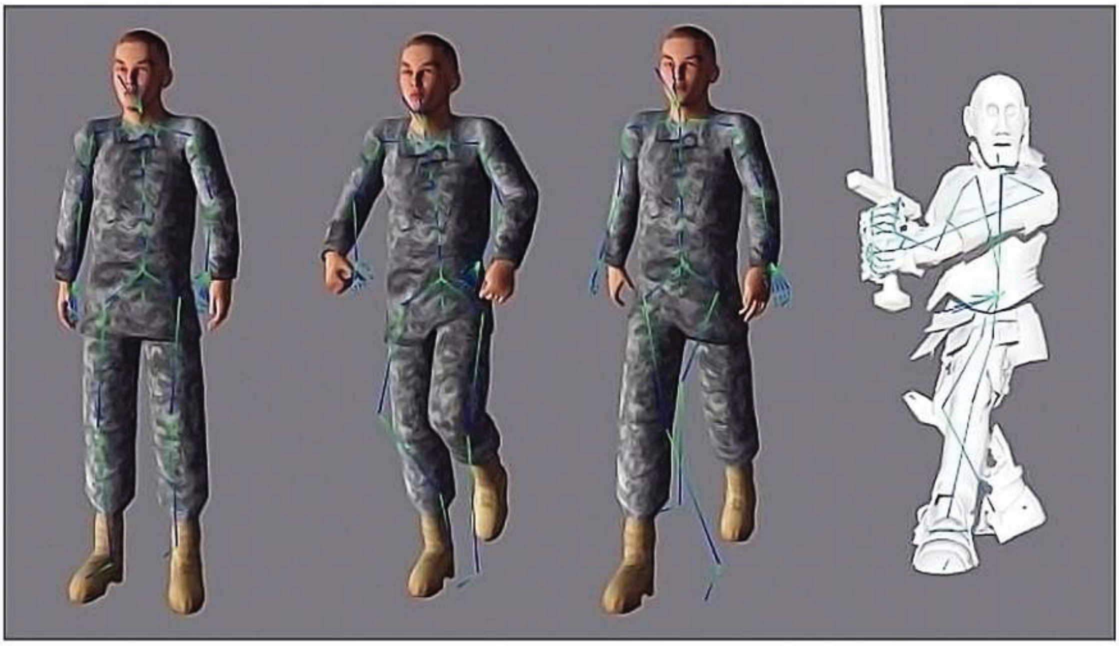


Figure 2.1: (Ahire et al., 2015) example of a 3D avatar with animations generated by blending or morph target systems that use controllers to define deformations and spatial position of the geometric mesh.

The avatar motions may be generated by controllers associated with the 3D geometric mesh using techniques such as Blend Shapes or Morph Targets as well as the tracking of geometric controllers based on a set of features and classifier cascade (Feng e Prabhakaran, 2016) (Ahire et al., 2015) (Kacorri, 2015). Some models can use notation based on sign languages, as Signwriting and HamNoSys, extending them for characteristics of manual signal elements, using terms such as symmetry, hand position, rotation and location, associated with values of movements or coordinates among other information (Connan e Moemedi, 2010) (Kaur e Singh, 2015) (Huang e Khan, 2016) as shown in Figure 2.2. These models make a parallel between sign language representation and their descriptive notations, where symbols are used to represent the terms and movements used by the speaker (Lombardo et al., 2011). Also, models like those mentioned typically do not cover facial expressions or give generic parameters defining an entire expression without covering the face as a set of independent landmarks.

McGowen e Geigel (2016) shown a process to use Facial Motion Capture to generate automatic new Blend Shapes. This process requires a full rig avatar with predefined parameters to be bound to controllers used on facial capture process. The input used by the authors was a video where facial landmarks were tracked and replaced by controllers in the 3D environment. Those controllers were rigged to a generic avatar which got the motion trajectory coordinates captured. Synthesis using 3D avatar usually is based on parameterized formal models. These models provide parameters and values being an essential resource to evaluate or increase an animation process. In the Soga et al. (2016) work a model for the automatic synthesis of choreographic dance through an avatar is presented. After that, a formal model was implemented, where parts of a human body received values based on dance movement parameters. This is similar to those parameters used in Sign Language synthesis where independent body parts receive coordinates or translation values. In general, animation of human characters requires complex calculations and high computational cost due to the large number of parameters required (Debuchi, 2017).

Human character structure can be represented as a hierarchical transformation of 18 body nodes with 31 Degrees of Freedom (DOF), where Lagrange equations enforce physical realism in synthesis animation (Jain et al., 2009). Based in the set of body nodes, gravitational forces can be deduced in addition with the kinetic energy of each body node. This model does not cover Facial structure, however.

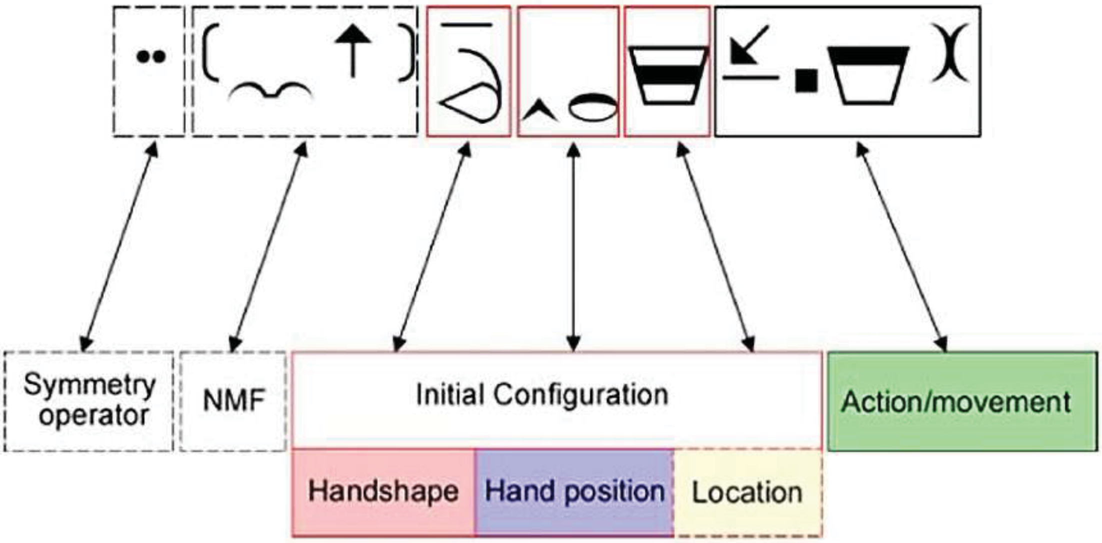


Figure 2.2: Parameters of a Sign, where we can perceive parameters such as symmetry and spatial values (Kaur e Singh, 2015). This type of parameterization is commonly extended to the whole body in formal models for signal languages.

Han (2015) develops a conversion system of human movements of a 2D input to a simplified 3D object output. This is achieved by an orthogonal connection analysis of the points and the reconstruction of features using 3D primitives. The major effort in this study was related to analysis and position estimation of each input joint, which gives in an orthographic plane of view a curve based structure creating a simplified avatar. Thus, it is possible to define landmarks in an avatar by their extremities, so that a 3D object can track the movements of a person connecting a reference point and its displacements by traced points of the joints.

An algorithm that extracts a 3D mesh from a representation of splines was implemented by Zou et al. (2015), where adjacent pairs of vertexes are traced to define the surface of a mesh structure allowing to perform the geometrical reconstruction, in a process called Skeletonization. This curve system is commonly used in algorithms which propose automatic rigging of 3D models, where the curves become controllers that usually assume the midpoint positions between joints of the model. This technique is also used to define 3D avatar features using the distance between landmarks and border limits of the object. Another method to construct an animated avatar is to build a rigging structure, connecting points in the geometric surface to controllers with defined Blend Shapes(Dailey et al., 2010). Blend Shapes is a prime animation method (McGowen e Geigel, 2016) where a rigging process is created associating vertexes weights to controllers. The geometric mesh is deformed using controllers with their associated weights in order to generate animation outputs.

Modulation in the mouth, eyebrows and other points of interest in the face can change the meaning of expressions in sign languages, besides details such as displacement and diffuse changes on virtual environment (Huenerfauth et al., 2011). Some key categories of expressions are identified as: *Question*, used when the sentence is interrogative, *Emphasis* used to highlight part of the sentence, *Emotion* as sadness and joy and *Continue* when the emitter have paused the message momentarily (Elons et al., 2014).

The base classes of emotional states used in interactive scenarios are the *Positive* (joy, surprise and excited emotions), *Neutral* (calm and relaxed expressions) and *Negative* (afraid, anger and sadness expressions) classes (Alkawaz e Basori, 2012) (Kocon, 2014). Ekman's model (Szwoch, 2015) identifies base expressions as *Anger*, *Fear*, *Sadness*, *Surprise*, and *Joy*. Other representations of emotions are identified as a combination and interpolation of basic expressions also called Plutchick's Wheel of Emotions (Figure 2.3). Shang et al. (2017) points out the main three feature descriptors to characterize human faces aspects: Action Unit (AU), commonly used to analyze facial expressions, Histogram of Oriented Gradient (HOG), commonly used for localizing face positions, and Felzenszwalb's HOG (FHOG), using principal component analysis. Concepts of some of these techniques will be used in experiments of this work in Chapter 3.

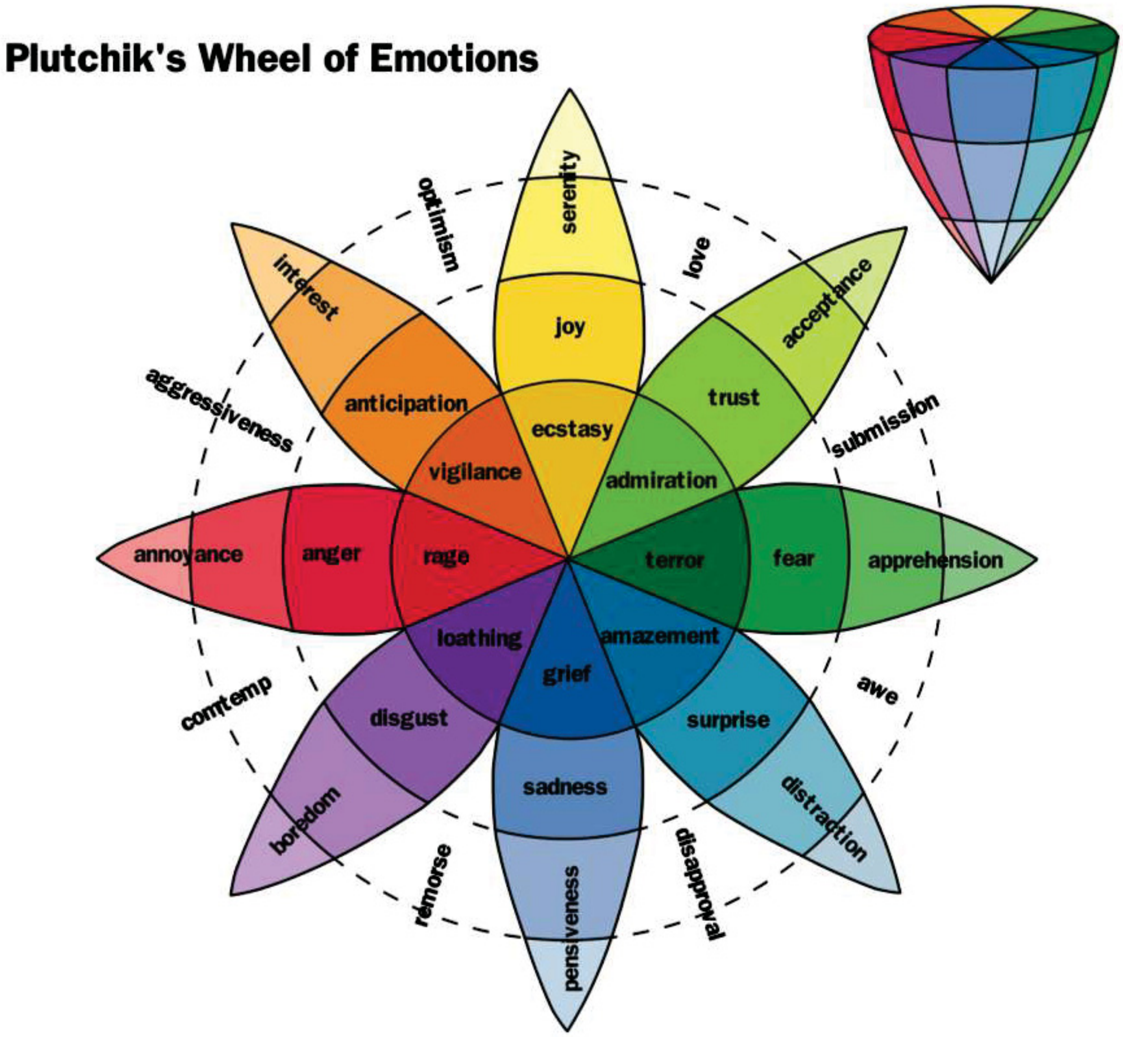


Figure 2.3: Plutchick's Wheel of Emotions. (Szwoch, 2015). This image relates the base expressions and interpolations joining intensities of more than one emotion. The model serves to identify possible main interpolations and their relation to base expressions.

Facial animations can be generated simulating the reactions of the muscles by anthropometric calculations using edge loops in two techniques: interpolation of shapes and deformation by controllers (Yamina e Farida, 2012) (Balci, 2004). An animation that uses shape interpolation involves the construction of a key series where displacements on geometric mesh are calculated in order to represent a spread deformation based on weight values for each vertex. Using controllers for facial deformations the visual changes of the geometric mesh are reproduced repositioning an array of vertexes by a spline controlled by a dummy primitive representing the sequence of vertices that follow the deformation by varying its intensity by proximity.

A Sparse Representation Classifier algorithm (SRC) can point landmarks like eyes and cheek from a depth channel (Lv et al., 2015). These process can be achieved using four binary masks based on distortions of expressions on a gray level image: rigid region (nose), semi-rigid regions (eye-forehead and cheek) and non-rigid region (mouth). Classification algorithms applied to faces can define areas with points where distortions are more perceptible. Facial Action Parameters (FAPs) can be used for emotion recognition and synthesis proposing methods based on Hidden Markov Models (HMM) for facial landmark cloud extraction (Zhao et al., 2015). These keys have the same architecture used in the sign language formal models (Kacorri, 2013), where keys and parameters are set to classify an element, like facial expression or emotion.

One possible approach to optimize the adaptation of the parameters contained in these 3D formal models is to modularize the models in order to apply local changes requiring fewer resources for each modification, instead of controlling the global model as an unique element. Krispel et al. (2018) presented a novel scheme to define varying states of a 3D model in a decoupled, declarative manner using pattern matching, with modular concepts applied to 3D objects as humanoid avatars but limited to the X3DOM format and Java scripts, which guarantees good portability but limits on computational cost.

A popular method to define facial landmarks and motion information is Motion Capture, where actors receive facial rig, usually with the use of electronic points or special markings, in order to capture coordinates of motion controllers. After that, this information can be used to animate a digital avatar (Cantwell et al., 2016). The tracked dots and shapes rotation are the base for blending shapes animation and can be used to synthesize signs. Aitpayev et al. (2016) shown a model for Static Hand Posture Classification, that uses image processing methods that resemble motion capture, based on tracking from a video input. Their method defines the corporal structure in six independent regions: the head, shoulders, neck, torso, hands, and belly. The region of the head is considered static, serving only as a reference to facilitate the tracking of the hands, the main focus in their research. The process then consists of observing and classifying shape, movement, and placement of hands.

The face can be decomposed into specific and independent regions handling the classification of key points for groups (Lemaire et al., 2011). These models are frequently used in Facial Expressions Recognition (FER) (Happy e Routray, 2015), where mathematical spatial models can define, for example, a triangle for eyes and nose, that can be applied to the synthesis process as well. MPEG-4 Facial Points is a broadly used standard set of points of interest in the face (Bouzid et al., 2013) (Balci, 2004). It has 84 points mapped on a model with a neutral expression, including areas such as tongue, lips, teeth, nose, and eyes, with points distributed along the perimeter of these regions, particularly at the corners. Talbi et al. (2017) uses the same parameters to define a segmentation model of the face surface in triangles for generation of automatic facial expressions, where vertices are repositioned simulating a facial expression in a 2D image. This model can be used as a base for setting facial controllers associated with a 3D geometric mesh, and also supports the use of additional features, such as variations in diffuse or bump textures (Jan e Meng, 2015), as well in albedo color representation of the face (Alkawaz e Basori, 2012).

For the purposes of this research, a well-defined set of landmarks should be used, once the proposed methods have the main objective of optimizing an emotion synthesis animation. Points in avatar face need to be mapped and temporally analyzed using methods for geometric simplification. Table 2.1 shows a linear ordered list of the works presented in this section and the relation with this thesis.

Table 2.1: Main Facial Parameters Bibliography References

Author	Year	Consideration
Huenerfauth, Matt et al.	2011	Define facial expression parameters .
Yamina, B. et al.	2012	Presented MPEG-4 facial model.
Bouzid, Y et al.	2013	Shows MPEG-4 parametric use.
Kacorri, Hernisa et al.	2013	Compare facial models for 3D avatars.
Kocon, M. et al.	2014	Define Expressions for Sign Languages.
Elons, A.S. et al.	2014	Define Facial Parameters for Arabic Sign Language.
Bento, J. et al.	2014	Define use of avatar for Portuguese Sign Language.
Mingliang Xue et al.	2014	Presents Facial Parameters based on Depth inputs.
Tangsuksant, W. et al.	2014	Define parameters for American Sign Language.
Kaur, S. et al.	2015	Define use of avatar for Indian Sign language.
Changqing Zou et al.	2015	Shows Skeletonization process for avatar animation.
Shiwen Lv et al.	2015	Presents an Local Region-based facial.
Yong Zhao et al.	2015	Extract a cloud of points from human face.
Jan, A. et al.	2015	Extract facial pattern using texture information.
Szwoch, W. et al.	2015	Presents a Model for Emotions.
Soga, A. et al.	2016	Developed a Body-part Motion Synthesis System.
Aitpayev, K. et al.	2016	Shown an region based method for Sign Language hand recognition.
Cantwell, B. et al.	2016	Investigate Motion capture for facial landmarks.
McGowen, V. et al.	2016	implements a Blend Shape for Motion Capture.
I. Talbi et al.	2017	Defines a polygonal facial model for generating facial expressions..
Debuchi, R.	2017	Proposes an easy 3D character animation editor.
Shang, Z.	2017	Evaluate Facial Expression Recognition for interaction with 3D avatar.
Basawapatna, A. et al.	2018	Reinforces the importance of 3D avatars in educational environments.
Wauck, H. et al.	2018	Analyze of the Avatar Self-Similarity effect on players.
Polys, N. et al.	2018	Reinforces The value of 3D models and immersive technology.
Hada, M. et al.	2018	Evaluates how 3D Avatar emotion affects their facial identification.
Krispel, U. et al.	2018	Present a novel scheme for varying of 3D model states.

2.2 4D AND SPATIO-TEMPORAL DATA

In order to solve problems such as facial expression optimization, even in a 2D scenario, only static information such as point-of-interest coordinates in a single state is not enough (Fang et al., 2011). The sequential information can be used in the process of locating a particular landmark by comparison of data such as motion, velocity or displacement. Most approaches use spatio-temporal information to classify facial data (Le et al., 2011) (Yan et al., 2008).

Visualization techniques of four-dimensional geometrical shapes rely on limited projections into lower dimensions, often hindering the viewer's ability to grasp the complete structure, or to access the spatial structure with a natural 3D perspective (Li et al., 2015). In certain approaches, a modeling of 4D data is not specifically related to a larger number of dimensions but to the use of elements of virtual reality as augmented reality (Flotyński e Sobociński (2018)). In (Serban, 2011) a mathematical model is presented to calculate spatio-temporal where, based on a orthogonal matrix, is considered the number of observations over time and the number of points of interest. V is the TxT matrix and Ul Vl is the empirical orthogonal function (or the midpoints of correspondence between the l^{th} values). According to Ngueyep (2014), one approach to model points within a defined space is calculate the distance between points (si, ji) and its distance in time (ti - tj) between the actual values (a1 ... an) being validated by a covariance function in a positive space.

In this case, a 3D object observed in a temporal slice, also called 4D, can be observed as an geometric object represented continuously as a single mesh representing the trajectory of all vertices. A Tesseract can show how this data should be represented in a spatial plane, shown in Figure 2.4, where we can observe in each image the increment of a dimension with the red line representing the direction of the depth axis. The first image has an edge bounded by two vertices in one dimension. Extending to two dimensions we can represent objects in a Cartesian plane, having coordinates for height and width observed by the square with four vertices and four edges. The same object can be extended to 3 dimensions by adding a depth, o "z" axis, so it is possible to observe a cube. The fourth dimension is the representation of two 3D cubes, with their connected vertices, where we can consider as the same 3D cube observed in two simultaneous moments, the new dimension being the temporal observation.

Different mathematical models for temporal data representation is found in the literature (Martin Erwig e Güting, 1998) (Lee et al., 2016) (Sikdar, 2017) (Yu e Poger, 2017) (Suheryadi e Nugroho, 2016). The general definition commonly used for 4D data is for an object observed temporarily in a defined space. The spatial data type, as well as the temporal environment, define the specificity's of the models developed.

Furthermore, the term 4D is used to represent, in some cases, augmented reality features, the specific representation of lighting and rendering in computer graphics, or geometric models that extend traditional 3D models by constructing more complex surfaces. In certain fields of study, the term 4D is also used for data dimensionality, for example in geographic or analytical data.

For Erwig et al. (1998), Space-Time Objects are the union of a point set in a Euclidean space and the temporal region observed along its continuous motion. It should be considered beyond the position in the space of a geometric object its extension, or the region of the movement and its spatial alteration in the observed window. For the authors, three fundamental abstractions for the representation of temporal objects are found: a point describes an entity in which location, is relevant. A curve between points in space, usually represented as a polyline or a spatial trajectory, describes a movement through space or spatial connections. A region is an abstraction

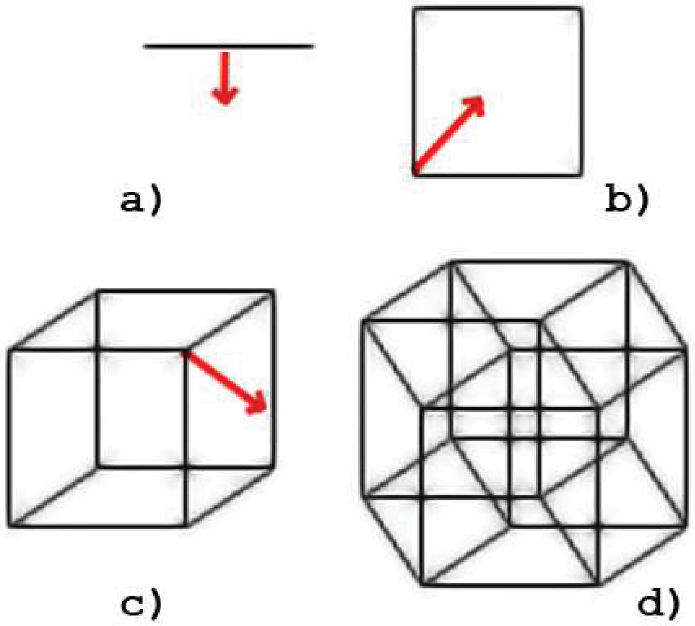


Figure 2.4: 4D object representation by a Tesseract object. Following the image: a) is an spline where two vertexes is connected by an edge, b) is an 2D representation by an square with coordinates x and y, c) is an 3D cube with depth dimension z and d) is an 4D Tesseract where two 3D cubes are connected by additional edges for all extremities.

for an object where its extension is relevant, for example, a data tree structure, or geometrically, a sequence of vertices connected by a synthesis context.

According to Oh et al. (2011) 4D is a kind of marketing term, considering that there is no standard rule or defined representation, for that, being used in different contexts. In movies the term is used when features using sensations involving smell or physical tremors are added to enhance the audio-visual experience (Oh et al., 2011) or even adding interaction in 3D movies (Lee et al., 2016). For Juanes-Méndez et al. (2016) the term 4D represents the union of a 2D image to 3D object information, merged using augmented reality, in order to increase academic applications.

Gonçalves (2016) uses the 4D term to represent the motion visualization in a sequential 2D input. Using voxel extraction from a video, the author compares the structural difference in order to find movement points in a temporal window. In that work the term 4D refers to dynamic points analyzed temporarily in a known space, resembling the concepts presented by Erwig et al. (1998) and Xu et al. (2017), adding a region view through voxels.

The term 4D can be used to represent a feature applied to a geometric object such as light rendering on dynamic faces or non-polygonal geometric surfaces (Chu et al., 2009). The term Spatio-Temporal is also used to represent different types of data, such as an extended Database (Martin Erwig e Güting, 1998) where the fields of a Data Table can be increased or decreased considering the concept of Data Region, as well as visual geographic data (maps or terrains) observed over a period of time (Sheidin et al., 2017).

Also using the term space-time to represent data, Sikdar (2017) shown a mathematical model to evaluate the spatial correlation and temporal correlation on geographical points. The authors propose a method to identify the existence of similar behavior between the use of gas in houses of the same geographical neighborhood. In Yu e Poger (2017) research, the authors use the term Spatio-Temporal to represent additional information that considers weight, or relevance, in fields of a Database based on analysis context.

For the purpose of this thesis, the term 4D is seen as a temporally observed geometric point. The idea of Spatio-Temporal Interest Point (STIP) occurs in points where their properties can be spatially relevant with a distinct location in the time corresponding to moments with inconstant movement compared to a space-time neighbor (Laptev e Lindeberg, 2003). The author represents space-time points of interest in a 3D environment where the original signal is represented by a threshold surface while the detected interest points are presented by 3D objects as shown in Figure 2.5:

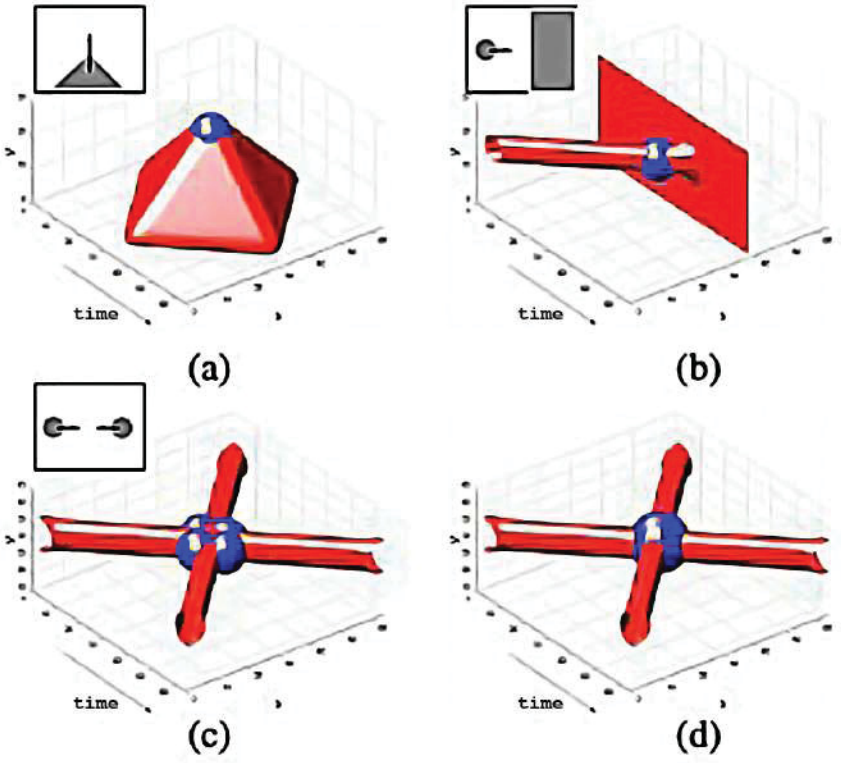


Figure 2.5: STIP (Spatio Temporal Interest Point) detection and rendered in 3D environment. a)Moving Corner; (b) A merge of a ball and a wall; (c) Collision of two balls; (d) the same as in (c) but using different scales; (Laptev e Lindeberg, 2003).

Figure 2.5 was obtained through methods to measure the spatio-temporal scale selection using the spatio-temporal Gaussian Blob algorithm and temporal variance combined with the Harris Interest Point operator. Then the authors present methods for classifying events using K-Means clustering of interest points. In order to define a model for a temporal object it is necessary to consider representations that are computationally traceable, that is, for an arbitrary temporal function $f \in \phi(\alpha)$ it is possible to determine the value of f at any time of the domain. In addition, the domain of ϕ is restricted to finite points or intervals in the defined time window (Martin Erwig e Güting, 1998).

(Martin Erwig e Güting, 1998) define a Temporal Object ω, as a set of disjoint and non-adjacent intervals. Φ is the spatial domain and τ is an type constructor which transform any given data type α into a type $\tau(\alpha)$ witch semantics $\tau(\alpha) = time \rightarrow \alpha$. In Erwig et al. (1998), the same author define a representation for spatio-temporal point ($\tau(point)$) and spatio-temporal region ($\tau(region)$). Assuming that space and time are continuous, a value of type $\tau(point)$ describes a function for a time position, represented by a curve in a three-dimensional space (x, y, t). A value of type $\tau(region)$ is a set of volumes in 3D space, representing a region of motion.

A type $\tau(point)$ and a type $\tau(region)$ relates vertexes in a geometric mesh relating a point in the known space and a volume extracted from the motion region. This concept can be applied to facial landmarks since their trajectory can also be converted to region volume using the same methods.

Suheryadi e Nugroho (2016) used Motion Patterns in Spatio-Temporal domain for a large range in order to detect moving objects. The authors used a video input where each frame was separated into horizontal and vertical slices by rearranging them into a spatio-temporal representation by 3D tensor set. From the vertical and horizontal slices of each frame, pattern recognition algorithms were applied considering points of intersection in the scene and eliminating the environment noise. In their work the Spatio Temporal term was used to represent the environment itself in sequential image slices.

According to Maniadakis et al. (2017), emotions affect the human perception of time, seeming to accelerate when the feeling is positive, as joy, and the opposite to negative ones as sadness and bored. In order to identify the relation of Emotions in the temporal perception, the authors used Radial Basis Function Neural Net with the level of positive and negative valence (happiness and sadness), as well the level of positive/high arousal, followed by the level of negative/low arousal. The authors further identified that the representation of emotions and their relation to temporal perception is also affected by the age of the observed person, where young people tending to be more expressive while older people often express their emotions for a longer time.

Similarly, Xu et al. (2017) shown methods for recognizing human activities in a video using Hierarchical Spatio-Temporal Model. The difference between a space-time action classification and common pattern recognition is that in the temporal domain, there exist strong time sequence relations between neighboring frames. The authors divide two categories for the recognition of patterns in human activities: temporal dependencies and spatial dependencies.

In the first case, by analyzing the activity of an individual in a time segment, or a sequence of frames, it is possible to identify, for example, if a person is drinking something or filling a glass, tracing the movements of his arm. By the other side, in order to identify a spatial dependency it is necessary to trace elements that interact with the individual in order to recognize, for example, if one person is hugging another or pushing. In order to point out temporal dependencies, algorithms such as Hidden Markov Models or Conditional random Fields are used to track patterns in consecutive frames. Spatial relations require the pose estimation methods in order to find parallel interaction.

Figure 2.6 presents pattern recognition outputs for dynamic points of interest that consider Spatio-Temporal relationships where, after identifying individual Spatio Temporal Interest Points, interaction of two or more points is considered in order to recognize a more complex activity (Xu et al., 2017).

The definition of a space-time point, in this case, requires the context of an action between more than one moving point, fitting the previously defined spatio-temporal region concept.

Spatio temporal information can be extracted using 2D image sequences reconstructing a 3D volume model (Yan et al., 2008) or comparing the differences between the mesh of a processed image sequences (Tim et al., 2014). This second method extracts volume information from frames where people performed specific actions such as walking or running. The authors use an information unit called Atoms, which consists of a temporal sequence of image slices representing movement in the horizontal direction with variation in gray tones of the pixels. Another technique uses a comparison of skeleton curves into 3D objects or Free Form Deformation (FFD) where two frames are compared and the movement of each vertex is interpolated to neighbor landmarks

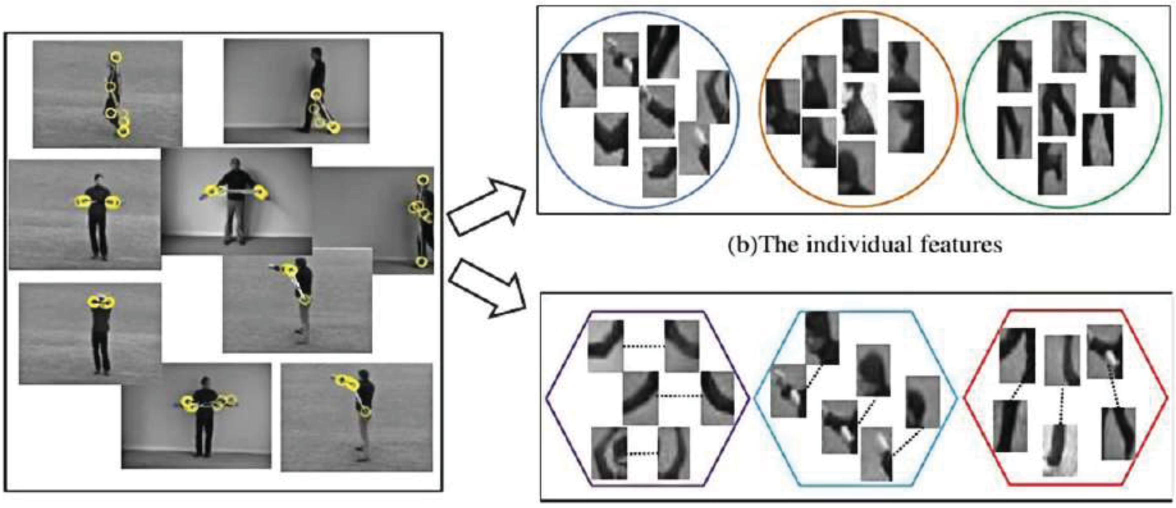


Figure 2.6: Recognition of Human Activity using Spatio-Temporal Model. a)Two learned high-level features; (b)Dynamic Features extracted from the input images; c) Each polygon represent the spatio-temporal frames, using relation between individual features; (Xu et al., 2017). The circles represent the tracked parameters and the hexagons show the temporal motions identified by the spatial parameters during the sequence of images.

(Zhang et al., 2014) (Sandbach et al., 2011). 3D volumetric mesh extracted to 2D images can have their surface curves compared and, for those discrepancies in depth, its possible to implement a classification system of facial expressions (Le et al., 2011).

Inverse Distance Weighted Interpolation Method (IDW) can be used to achieve spatial interpolation (Hongjian et al., 2016). So, with the geometric coordinate captured, spatio-temporal data can be defined by calculating the distance from each detection point to the insertion point firstly, based on the polynomial interpolation of the tabulated data.

Facial Expressions Recognition can consider spatio-temporal information (Mahmoud et al., 2014). Sun e Yin (2008) compared frames related to an expression obtained through a 3D face captured by a scanner hardware where are extracted 83 landmarks and then adjusted to a generic template and applied Hidden Markov Model for the recognition of facial expressions. Using a similar process it is possible to extract depth points on the object and classifies the facial expression of the object using spatio-temporal data extracted (Xue et al., 2015) (Xue et al., 2014).

There are works that specify the 4D data not as an temporal information, but as a 3D sequential abstract features like weight influence of vertexes or impact area in simulations such volume or illumination. Chu et al. (2009) shown an architecture for viewing 2-manifold objects and 3-manifolds. The main concept of the architecture presented is the render of a tetrahedron using hyper-volumes or voxels, to represent the surface or layer of elements, unlike the traditional use of the polygonal system. The challenge then is the correct representation of illumination incidence as shown in (Figure 2.7).

There are 3D-face databases with sequential images captured by devices such Microsoft Kinect's with provides deep channel (Hong et al., 2015) where is possible calculate temporal information using the position of the vertexes and sequential displacements (Cao et al., 2014). According to Zhu et al. (2015), Spatio-Temporal Interest Point (STIP) can be obtained using the algorithm Harris3D, which locates points of interest based on a gradient matrix in sequential images with the Gaussian filter. This technique finds spatio-temporal volume where gradients variations appear with less notable sharpness, showed in (Figure 2.8) followed by a Hessian detector which measures the strength of each landmark.

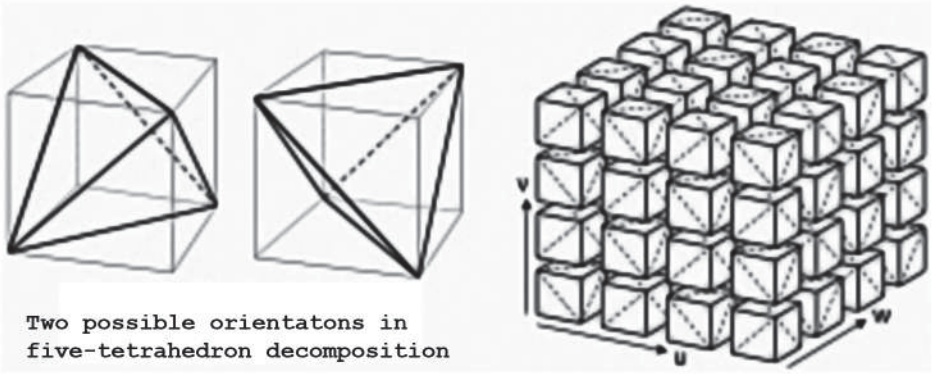


Figure 2.7: Thetraedron decomposed surface. The normal or bump channel have no direction information to use once this are not an 2-manifold object (Chu et al., 2009).

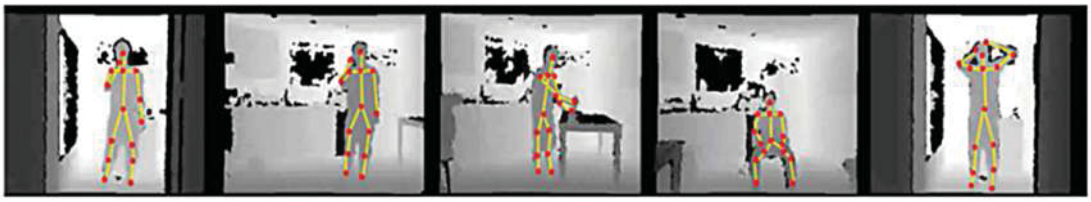


Figure 2.8: Harris Corner 3D used to find volume information comparing textures information and tracking for an human armature. (Zhu et al., 2015).

Vemulapalli et al. are presented in 2014 a model for the recognition of human actions considering spatio-temporal information comparing key points in a 3D skeleton. The approach of this work was to show the relationship of body parts in an action using a skeleton of 3D points. According to the author, human actions can be represented by curves, using modeling of special Euclidean group SE(3), widely used in works where movements such as rotation are relevant and calculations for rigid body kinematics will be needed. The differential of this research was to analyze the body and its actions observing pieces not directly connected but objectively relevant in a particular action as hands-on clapping action. This method can be applied to locate non-human objects as well (Amini e McGuffin, 2014). Huang (2017) defines a Temporal Fitting formula for recognition of human actions using concepts of motion capture and 3D mesh overlaid on 2D input.

According to (Kai et al., 2014) the landmarks topology in the body is in a theoretical limit, where studies focused on consolidating the captured geometry topology requires more attention. The author presents a formula which motion data can be represented starting from the initial values of the point, based on the hierarchy trees.

$$F_i = (p_0(t), q_0(t), q_1(t), q_2(t)...., q_n(t))$$

where the first points p and q represent the position and orientation of root nodes, the following points follow the tree hierarchy.

These models can be used to represent temporal facial landmarks in an avatar. Still, with the modeled and extracted data, geometric or animation optimization can be built considering their influence in the course of an animation. Table 2.2 shown the main references of this section and considerations for this work.

Table 2.2: 4D data Bibliography References

Author	Year	Consideration
Erwig, M. et al.	1998	Presented Spatio-Temporal Data Models and a Comparison of Their Representations.
Erwig, M. et al.	1998	Discrete Modeling of Spatio-temporal Data Types.
Laptev, I. et al.	2003	Define a model for Space-Time Interest Points.
Chu, A. et al.	2009	Proposes a GPU-based Architecture for Interactive 4D Visualization.
Oh, E. et al.	2011	Analyzes how 4D Effects affect user experience.
Serban, N. et al.	2011	Presented Spatio-temporal modelling.
Sandbach, G. et al.	2011	Define temporal models for 3D face.
Vuong Le et al.	2011	Use spatio-temporal in FER.
Amini, F et al.	2014	Implements an 3D visualization for movement data.
Tim, S.C.W. et al.	2014	Presents Temporal Action Recognition by 3D patches.
Vemulapalli, R. et al.	2014	Implements Action Recognition using 3d skelleton.
Hui Zhang et al.	2014	Proposes an method for curve handles in 4D.
Hong, Richang et al.	2015	Presents RGB-D uses.
Zhu, Yu et al.	2015	Implements Depth-Based Action Recognition
Mingliang Xue et al.	2015	Propose an 4D facial recognition method.
Li, N. et al.	2016	Present an mobile visualization for 4D objects.
Hongjian, W. et al.	2016	Use IDW to define spatio-temporal data.
Jaebong, L. et al.	2016	Presents an Interactive Motion Effects Design for a Moving Object in 4D Films.
Suheryadi, A. et al.	2016	Apply Spatio-temporal analysis for moving object detection under complex environment.
Mendez, J. et al.	2016	4D Environment for Learning in the Human Anatomy Field.
Sheidin, J. et al.	2017	Visualizing Spatial-Temporal Data.
XU, W. et al.	2017	Presents A Hierarchical Spatio-Temporal Model for Human Activity Recognition.
Songmei, Y. et al.	2017	Uses Temporal Weighted Data Model in education environment
Sikdar, B. et al.	2017	Proposes Spatio-Temporal Correlations in Cyber-Physical Systems.
Huang, Y. et al.	2018	Defines a Temporal Fitting formulae.
Flotynski, J. et al.	2018	Presents an Semantic 4-dimensionai modeling.

2.3 3D MESH AND ANIMATION OPTIMIZATION

3D Mesh optimization helps to understand the region simplification of a model (Ko e Choy, 2002) and can be applied to avatar face in order to analyze how the geometric structure or an animation process can be temporally optimized or interpolated. In 2014, Ng e Low shown a mesh simplification method is presented based on the Euclidean distance of the edges for each triangle where the size of the edge and the distance between spatial coordinates values of adjacent triangles are used to calculate polygons reduction in the surface normalizing the adjacent vertex of the triangle to the perpendicular edge generating the process (Figure 2.9):

This technique, called Half-Edge contraption, provides a simple solution for reduction and simplification of polygonal structures. This process has a direct parallel to the animation

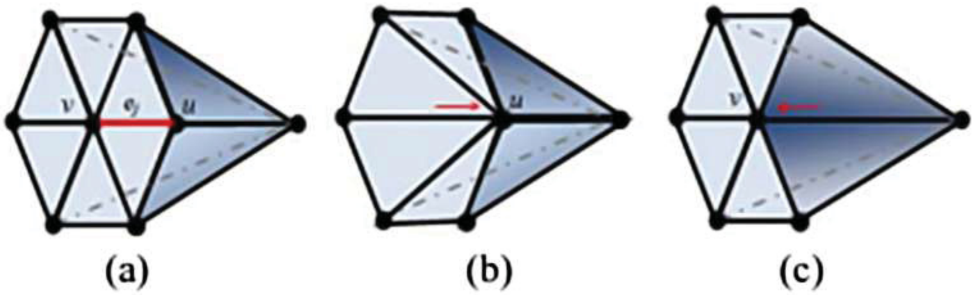


Figure 2.9: Mesh Simplification Process: The centroid of neighborhood for a ring of vertexes are calculated and a merge of near edges occur following an displacement to center of mesh (Ng e Low, 2014).

process if vertexes have been observed as a key of Blend Shapes set. Blend-Shape is a technique used in computer graphics that consists of defining displacements in a geometric mesh based on a time interval, using 3D controllers and concepts of sequential animation. Cetinaslan e Orvalho (2018) presents the method that localizes the direct manipulation of blend-shape models for facial animation with a customized sketch-based interface. The authors argue that methods of direct manipulation of animation by blend-shape can result in unexpected deformations in the mesh, since they affect the structure globally, proposing a technique that uses geodesic circles for confining the edits to the local geometry (Figure 2.10). In this way, although the creation of animations by blend-shapes remains conceptually the same, the deformations have a more precise impact.

Baoming e Xuena (2013) present a 3D object simplification proposal, using an additional table to indexing edges, averaging distances between points. This table indexes the vertexes temporal coordinates organizing a list of structs pointing sequential edges. For mesh simplification authors extended the Garland algorithm which constricts pairs of vertexes (Figure 2.11).

So for simplification of the polygonal structure, values of the centroid faces are summed, as represented by Q(v) and are observed the area and weights (w) of the triangle while the original geometric shape is kept, also reducing breaks in the polygonal structure. Like the previous method, these can be compared to an animation optimization placing output centroid as expression landmark position and the simplified points as landmark temporal changes.

$$Q(v) = \sum_i w_i Q_i(v)$$

The process follows identifying edges that share triangular faces that are deleted creating a vertex in the empty space linking the neighboring edges. The basis for the geometrical simplification consists in keep the shape of the object similar to the original by reducing the triangles where the surface remains flat (Park et al., 2002). For that, some techniques can be used as a Polygonal Approximation of Continuous Curve and Polyhedral Approximation of Continuous Surface (Figure 2.12).

These models for mesh simplification applied to a virtual interpreter for sign language systems already reduce the computational cost, however, the methods presented can be adapted to fit an analyze and displacement reduce of the Blend Shape animation process. Added to facial models, its possible obtain a parametrized control of the geometric facial surface. In addition to simplification using geometric meshes, it is possible to reconstruct a facial structure based on Voxels, as presented by Li et al. (2018) which uses volumetric data structure and Deep Neural Network to predict the variation of human faces.

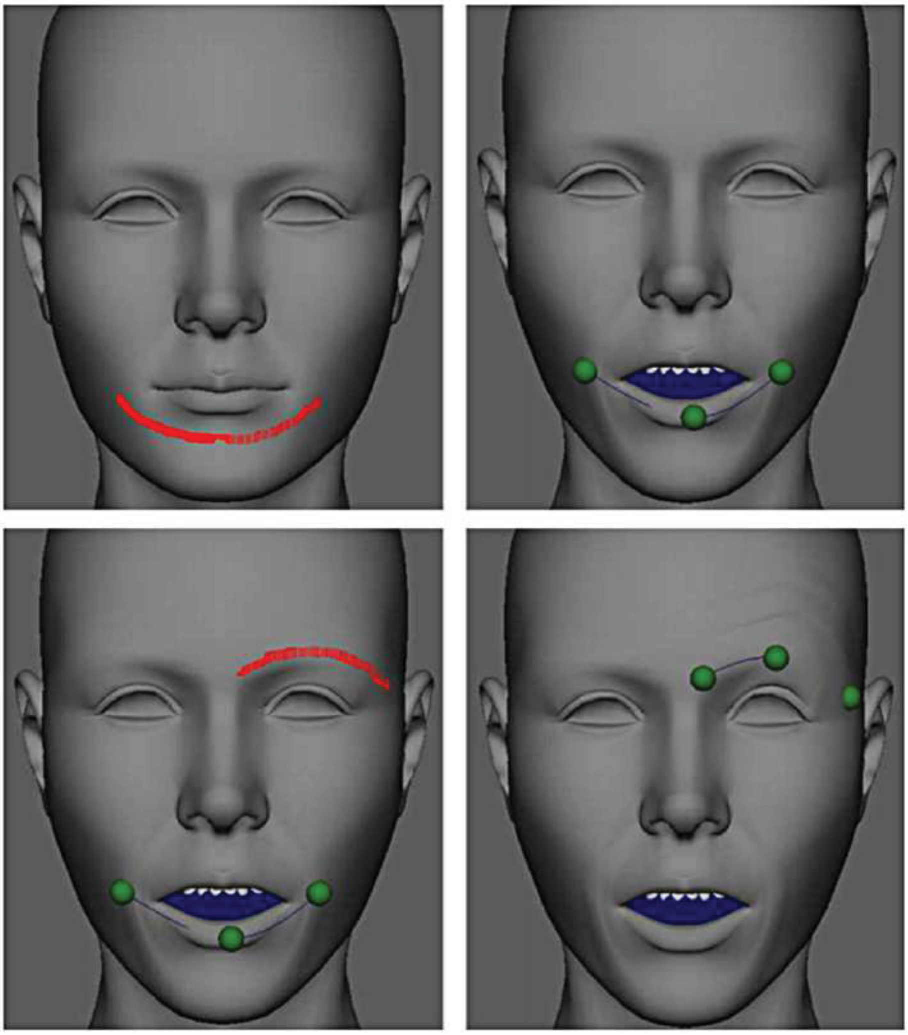


Figure 2.10: (Cetinaslan e Orvalho, 2018) Blend-Shape animation method.

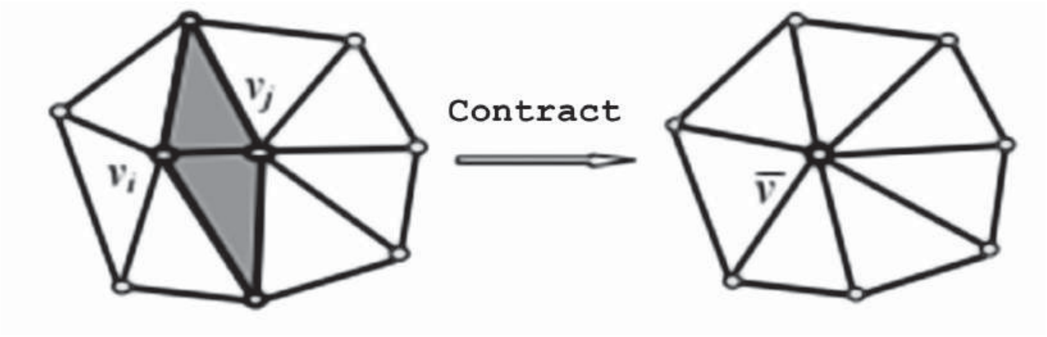


Figure 2.11: Baoming's Mesh Simplification Method based on Garland algorithm where pairs of vertexes are merged (Baoming e Xuena, 2013).

An interesting feature used in mesh simplification can also be used in the animation optimization, which are Normals of the vertexes in the metric of geometric distortion (Jian et al., 2013). Normal field variation determines the orientation for landmarks, describing features of the mesh surface (Figure 2.13) which allows the application of algorithms like QSlim where the geometric mesh is reduced by eliminating small triangles in flat regions by their orientations. There are more specific techniques in which a geometric surface can be reconstructed, as well as

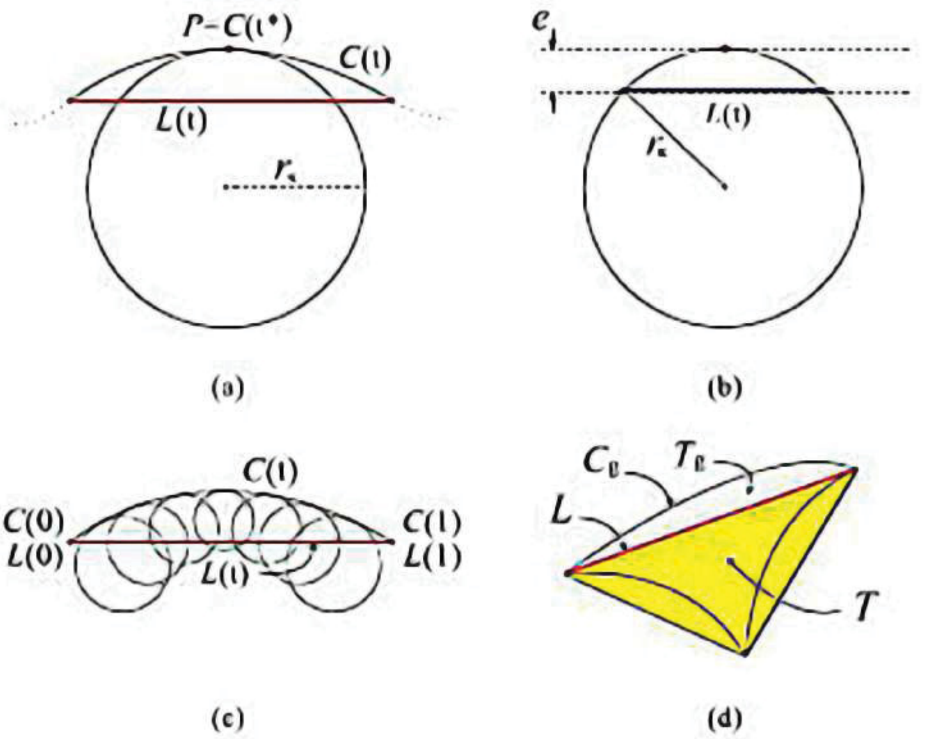


Figure 2.12: Surface simplification where a polyline segment excludes the corresponding curve: a) *P* be a point on a curve *C(t)= (x(t),y(t))*, which is abstracted by a edge *L*. b)The optimal tangential length at P is defined as the maximum possible length of the chord in the tangential circle when the maximum arc-edge distance is below *e* c)Tangential circles representation d)In 3D surfaces a triangular facet abstracts some portion of the surface, which is called the geodesic triangle (Park et al., 2002).

a curve, removing a single triangle, or the smallest number of control points (Methirumangalath et al., 2017).

Setty e Mudenagudi (2018) describes a technique of rebuilding damaged 3D meshes using point clouds method. A region of Interest (ROI)-based method is used to define which regions of the model can be filled based on surface analysis by related points. ROI can be applied in facial animations, defining independent areas for behavior analysis of vertices or polygons.

Algoritihms such Data Time Warping (DTW), Factor Analysis (FA) and Principal Components Analysis (PCA) can be used to understand landmarks correlation and your behavior sequentially. DTW is a distance measure that compares two-time series after optimally aligning them (Mueen e Keogh, 2016), which can be applied to facial landmark during an expression synthesis. PCA an FA can define facial landmarks expression and weight during the same synthesis, aligning a landmark series by their relevance. DTW process can be optimized and applied to activity recognition through data normalization and the use of the correct warping-window parameter. Some of those concepts will be deeper discussed in Chapter 5.

Animation interpolation can be observed in the scene observing the camera's point of view beyond the perspective and depth of field (Mori et al., 2015). According to the authors the greatest deficiencies of systems that use virtual environments are unrealistic or robotic animation, which may be increased by the user scene point of view. Considering the scene viewpoint, its possible optimize an animation through trajectory controller. To this can be calculated a B-Spline that directs the pivot points or joint control of an object in a frames sequence, thereby minimizing the expected trajectory synchronized with a new constrained spline (Schoch et al., 2014). This

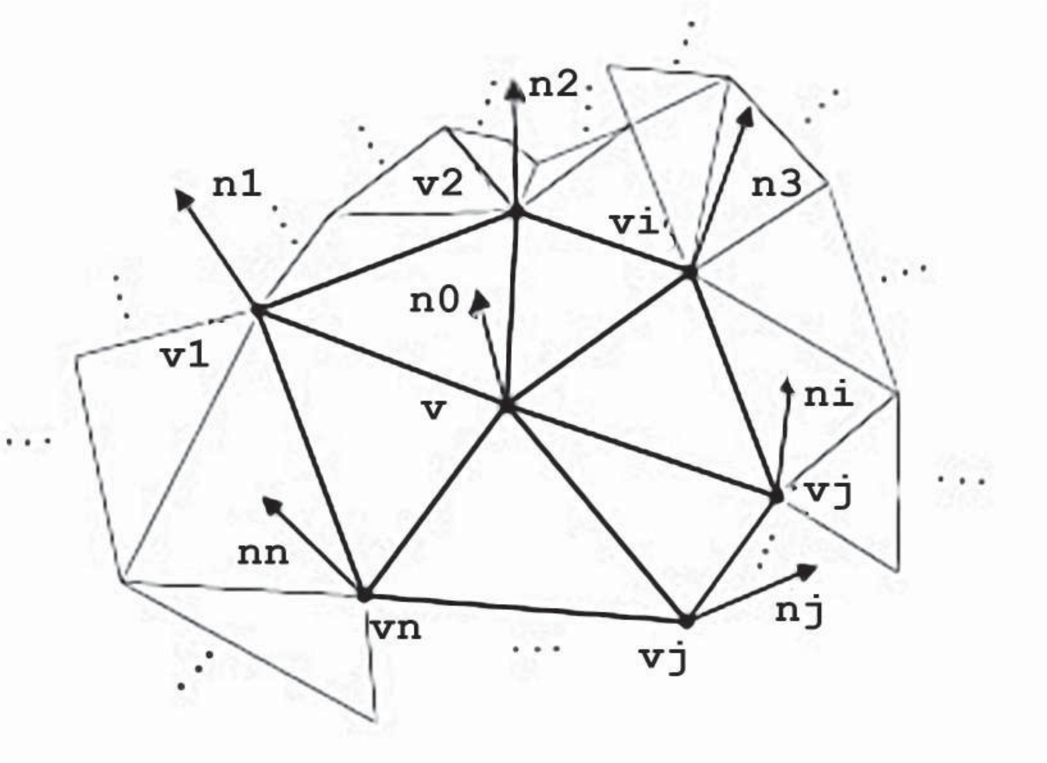


Figure 2.13: Normals of neighborhood applied to an landmark in geometric surface. Those directions can be used in mesh simplification (Jian et al., 2013).

can be applied in sign language synthesis if consider the points as joint parameters such mouth or cheek corners.

Similarly, it is common to use estimates of trajectories for the animation optimization for robotics and virtual reality simulations. The trajectory is a continuous set of curves representing the states and inputs to the system over the time interval (Johnson e Murphey, 2009). The trajectories can be drawn from position calculations obtained from points of interest in the 3D object, these extracted or identified from models such as MPEG-4. Another interesting way to identify points of interest on objects is extracting the skeleton of the geometric structure (Yamane e Goerner, 2014), which allows through sequential analysis, get the points of articulation, and thereby observe its trajectory (Figure 2.14).

Trajectories can be used very efficiently to analyze the behavior of an element in an interval of readings in a virtual or real environment. Vu et al. (2017) uses calculations for Trajectory Reconstruction in the context of pedestrian location. The concepts used in your research can be adapted to facial points in a 3D avatar if considers the environment as the geometric mesh and the tracked pedestrians as landmarks. A similar application applied to 3D meshes is presented by Pan et al. (2018), where, using Neural Networks, methods for reconstruction of UV coordinates (usually representing the opened 3D mesh in a 2D plane) are extended for the reconstruction of the geometric mesh. Idaka et al. (2017) uses local region calculations for pose estimation mapping a basketball court with edges and analyzing the trajectory of dynamic objects.

Geng et al. (2014) uses 3D trajectories to recognize Chinese Sign Language gestures using image processing in depth channel inputs obtained by the Microsoft Kinect device. From the hand tracing the author used classification algorithms such as Support Vector Machine (SVM) to identify, based on the 3D trajectories, the represented signal (Figure 2.15). Trajectories can be used for the sign language alphabet letters recognition using hand position (Sulfayanti et al.,

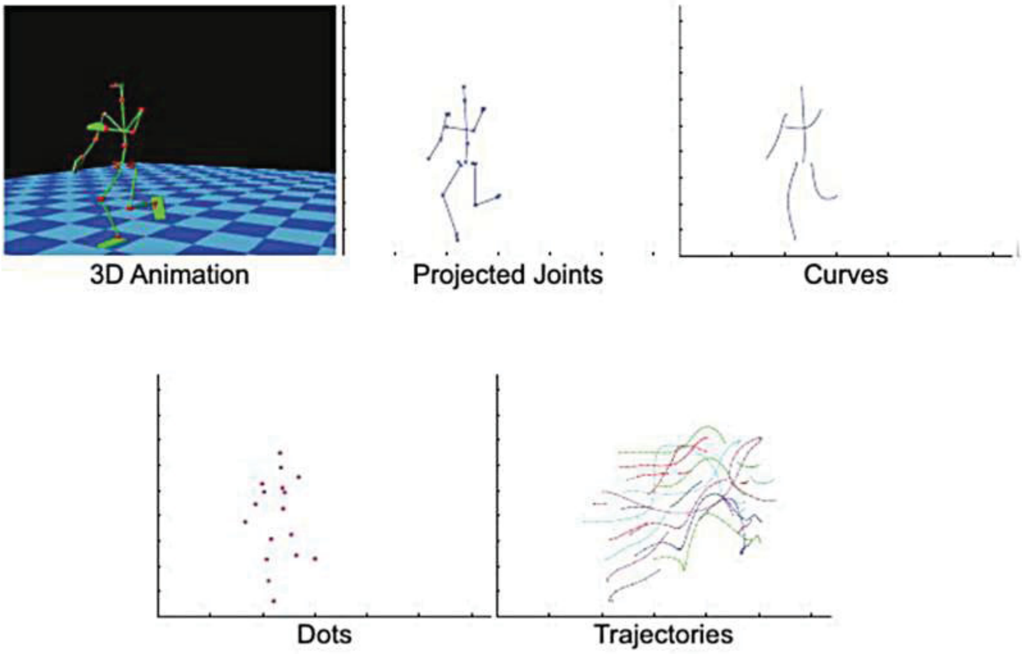


Figure 2.14: Trajectories from 3D animations, showing the path of landmarks extracted from the object (Yamane e Goerner, 2014).

2016). The author uses the depth channel to identify the hands and the position of the fingers and traces the position generating a trajectory, where it is possible, by calculating the Euclidean distance of specific points from the user's hand, to infer a letter being represented.

Gajalakshmi e Sharmila (2017) also uses SVM to recognize Sign Language in videos. The author uses training classes for some specific signals, or words, and trains repeated feature histograms patterns. One of the main challenges for hand signal recognition is the hand position, usually using techniques for aligning the input, making rotation, scale or translation irrelevant in the signal mapping. For facial synthesis, the rotation is not relevant since the emotion can be represented independently of the speaker's head position. There are, however, studies that argue that the position of the head can help in the classification of a questioning expression (Huenerfauth et al., 2011).

Oliveira et al. (2017) uses the PCA + k-NN algorithms to extract features from an image base referring to the alphabet in Irish Sign Language. According to the authors, techniques used for hand sign recognition include algorithms such as Hidden Markov Models, Points of Interest, Principal Component Analysis (PCA), Orientation Histograms and Kalman Filter, as well as classifiers such as CNN and k-NN. The PCA algorithm, in particular, can be used to classify the behavior of facial features by defining which landmarks have the greatest variance between tabulated reading values.

The main way to control animations in virtual environments is through data-driven, files or specifications that define elements such as controllers and paths with coordinates that tell how certain object or surface will behave in a sequence of actions. The own controllers can be optimized, as proposed by Lee et al. (2009) where using animation through parameterized data, the motion selection determines which controllers turn on and their positions and compared with the same actions, performed manually, the authors defined calculations to automatically select decisions for actions control and movements with lower computational cost compared to random or manuals selections.

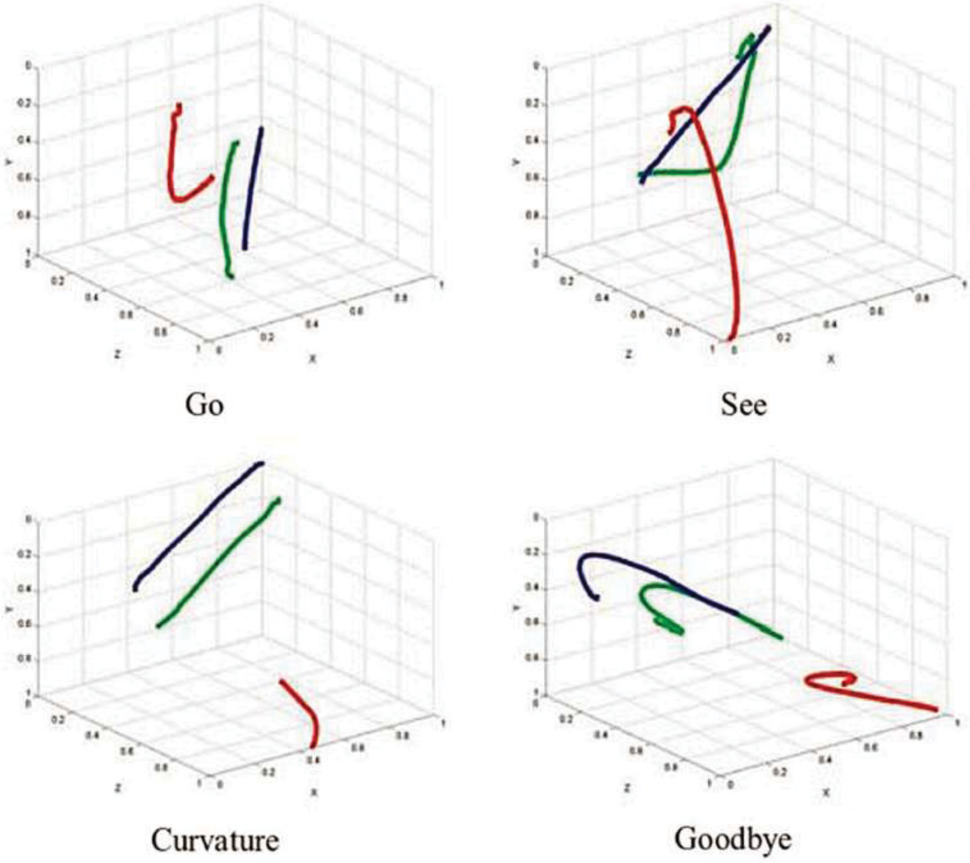


Figure 2.15: 3D Trajectories used for Chinese Signal Language recognition (Geng et al., 2014).

Agus et al. (2017) uses Data-Driven to analyze the behavior of users in exploring a virtual environment using measures on surfaces. Data-driven similar techniques can use other input models to generate for 3D applications, such as the Query-based composition of animations (Flotyński et al., 2018) or in coordinate analysis (Agus et al., 2017). In this method, a hierarchical parameter-based formatting is used to define specific joints associated with a previously developed rig, where motion sequences for a 3D avatar are pointed out. This template can be adapted to the Tags format used in XML as well as the signal language parameters.

Signal language parameters are usually registered in XML tag format, in models that follow the pattern similar to presented by Othman e Jemni (2017), as follows the code generated:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
   elementFormDefault="qualified"
   xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="points">
   <?xml-stylesheet  type='text/xsl'  href='form.xsl'?>
   <xs:complexType>
    <xs:sequence>
    <xs:sentence srclang="en" lang="asl" srcsentence="i have
       going to the conference for five years.">
    <xs:clause typeclause="none">
      <xs:token>I</token>
      <xs:token property="HABITUAL">GO</token>
      <xs:token>CONFERENCE</token>
      <xs:token compound="-">
         <xs:token>UP</token>
         <xs:token>TO</token>
         <xs:token>NOW</token>
      </xs:token>
      <xs:token numeral="yes">FIVE</token>
      <xs:token numeral="yes">YEAR</token>
    </xs:clause>
      </xs:sentence>
    </xs:sequence>
  </xs:complexType>
 </xs:element>
</xs:schema>
```

Where we have the output "$GO^{Habitual}$ CONFERENCE UP-TO-NOW FiveYear" which means, according to the authors, "I have been going to the conference for five years." in ASL (Othman e Jemni, 2017). This hierarchical organization in Sequential Tags can be applied to facial parameters including specific information such as coordinate or position of a landmark, or even values for regions of the face that assist in the segmented representation of facial expressions. Ji et al. (2016) uses an output method for manual Sign Language classification exporting a binary value for each letter in signal languages alphabet. Its model differs from previous ones that use descriptive tags, approaching closer to the coordinate valuation model working similar to an encryption record.

The use of data-driven and XML format for parameter registration will be explored in this work in order to identify ways to optimize the facial expressions animation using the BSL parameters associated with facial points from MPEG-4 model. Another aspect to be considered in optimization and realism of 3D animation is the Balance Controller of scene bodies (Jain et al., 2009), where methods for weight distribution can be applied, calculated by collision and weight projection (synchronized to surface vertexes) may addition force aiming balance of the movement.

Yet, using a joint armature or landmarks structure, it is possible to calculate the behavior of 3D avatar animations (Daoudi et al., 2018) or interpolate an animation through Inverse Kinematic algorithm (Ben Yahia e Jemni, 2013). The first step in this type of technique is to extract the points of interest that can be realized through a skeletonization process of geometrical surface or on formal models, which for human faces should consider edge loops or models based on muscle strain. Through this information a trajectory forecast is made based on the displacement of the points and influence among their neighborhoods (Figure 2.16), allowing not only set a more appropriate sequential deformation as built optimization strategies based on this information. Besides, that can be uses features for muscle control minimizing the change of joint torques, increasing realism to organic animations.

From the methods and concepts cited in this chapter will be presented the thesis proposal and preliminary practical experiments focused on facial parameters definition. Table 2.3 shown main references of this section and the relationship with this work.

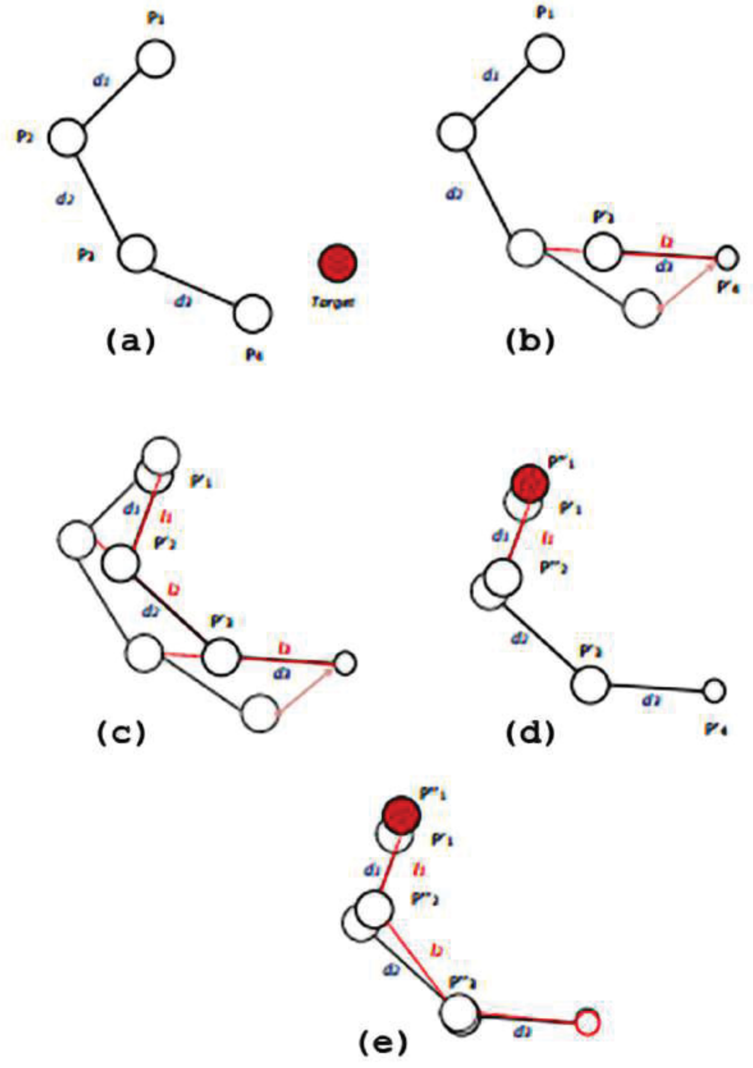


Figure 2.16: Animating 3D Avatars using Inverse Kinematics: In (a) predicted landmark displacement is calculated, in (b) and (c) sequential landmarks are aligned based on the target. In (d) and (e) all segment is reallocated aligned to a new target positioned on base joint (Ben Yahia e Jemni, 2013).

Table 2.3: Mesh and animation optimization Bibliography References

Author	Year	Consideration
Jain, Sumit et al.	2009	Presents an optimization Motion synthesis method.
T. D. Murphey et al.	2009	Proposes an trajectory synthesis from animation data.
Hou Baoming et al.	2013	Geometric Simplification using point weight.
Wang Jian et al.	2013	3D Mesh Simplification using Normal coordinates.
Ben Yahia et al.	2013	Uses Inverse Kinematics for sign avatar animation.
Kok-Why Ng et al.	2014	Presents an Geometric Edge Contraption method.
M. Schoch et al.	2014	Proposes an trajectory optimization method.
K. Yamane et al.	2014	Implements an trajectory skeletonization.
L. Geng et al.	2014	Uses 3D trajectory for hand gestures recognition.
Mori, Hiroshi et al.	2015	Proposes an animation method by viewpoint.
Mueen, A. et al.	2016	Present DTW optmization method for landmarks.
Ji, Y. et al.	2016	Exports a binary value to the manual alphabet in signal languages.
Sulfayanti. et al.	2016	Uses hand tracking for Sign Language recognition.
Agus, M. et al.	2017	Present Data-Driven Analysis of an 3D environment.
Idaka, Y. et al.	2017	Uses Local Region calculation for pose estimation.
A. Othman et al.	2017	Propose a XMl model for Sign Languages parameters.
Methirumangalath, S. et al.	2017	Proposes 3D Reconstruction method. approach.
Vu, H. et al.	2017	Builds a method for pedestrian Trajectory Reconstruction.
P. Gajalakshmi et al.	2017	Uses SVM for Sign Language signals recognition.
M. Oliveira et al.	2017	Applies the PCA and kNN algorithm to identify signs of the Irish Sign Language.
Flotynski, J. et al.	2018	Implements an Query-based composition.
Cetinaslan, O. et al.	2018	Present Blend-Shape 3D techniques for local mesh deformation.
Setty, S. et al.	2018	3D Mesh reconstruction based on point clouds.
Li, X. et al.	2018	Define a Volumetric model for 3D Facial reconstruction.
Pan, Z. et al.	2018	Presents an Automatic Re-topology for 3D Scanned Objects.
Daoudi, M. et al.	2018	Uses Trajectory Analysis to identify Human Actions.

3 PROPOSED SIGN LANGUAGE FACIAL MODEL BASED CENTROIDS

The main deficiencies and challenges observed on Chapter 2 concerning facial features are related to the definition of low-cost facial models with optimized controllers and methods. It's important to remark that BSL has no formal models for facial representations, which are indispensable for computational notation of a Sign Language.

One approach to understand the whole process of an expression animation is to analyze the geometry temporally, what makes 4D information an important feature in this scenario.

This work investigates facial animation for Sign Languages synthesis proposes methods for facial animation generation using 4D information in geometric surfaces. Initially, it was proposed a facial model that represents the face in a parameterized way, making possible to apply 4D modeling for observation of spatio-temporal facial landmark behavior.

In order to define the facial landmarks and interested regions, the Base Emotions Ekman's model was used (Szwoch, 2015) (Lyons et al., 1998) (Lundqvist, 1998). The enhanced model defined in this chapter allows the temporal analysis of the facial structure, represented by the 3D mesh, allowing a study of 4D data modeling. The next sections define the proposed facial model, the 3D avatar built for the experiments of this work, as well a Principal Component Analysis of the facial expressions in this context.

3.1 FACIAL MODEL AND REGIONS INFLUENCE

The objective of the model proposed in this chapter relies on a hierarchical modeling compatible with the models widely used in the computational representation of the manual parameters of sign languages, improving systems that do not use expressions or emotions on message transmission.

A 3D avatar was built with a geometry compatible with the facial datasets that requires a low poly mesh (a simple geometry with few polygons), and that can efficiently represent the base expressions defined by Plutchick's Wheel of Emotion (Figure 3.1). Both the geometric model and the parameters defined in the following subsections can be used to represent any facial expression using interpolation of facial landmarks displacement values.

The following experiments are divided into two stages: 1) Proposal and construction of the BSL Facial Model, applied in an avatar for emotional representation in sign language systems. 2) Principal Components Analysis of facial geometric surface in the expression representation context.

3.1.0.0 BSL Avatar Facial Model

A humanoid model was built with 598 facial polygons and a rigging supporting the aforementioned base expressions (Figure 3.1). The facial model MPEG-4 FP (Kacorri e Huenerfauth, 2014) was used as a reference in avatar modeling (Figure 3.2). Also, the most representative regions of the face were set: forehead, eyes, cheeks, nose, and mouth. The regions were defined on the basis of experiments shown by Obaid et al. (2010), describing main regions for facial deformations. Other studies using facial regions for the extraction or classification of characteristics defend the local division as a fundamental resource to better understand the behavior and relation between a geometric area or related spatial points (Lemaire et al., 2011) and (Lv et al., 2015).

The avatar was built with few polygons aiming to avoid high computational cost besides a greater flexibility when calculating displacements in the 3D mesh during the animations. In

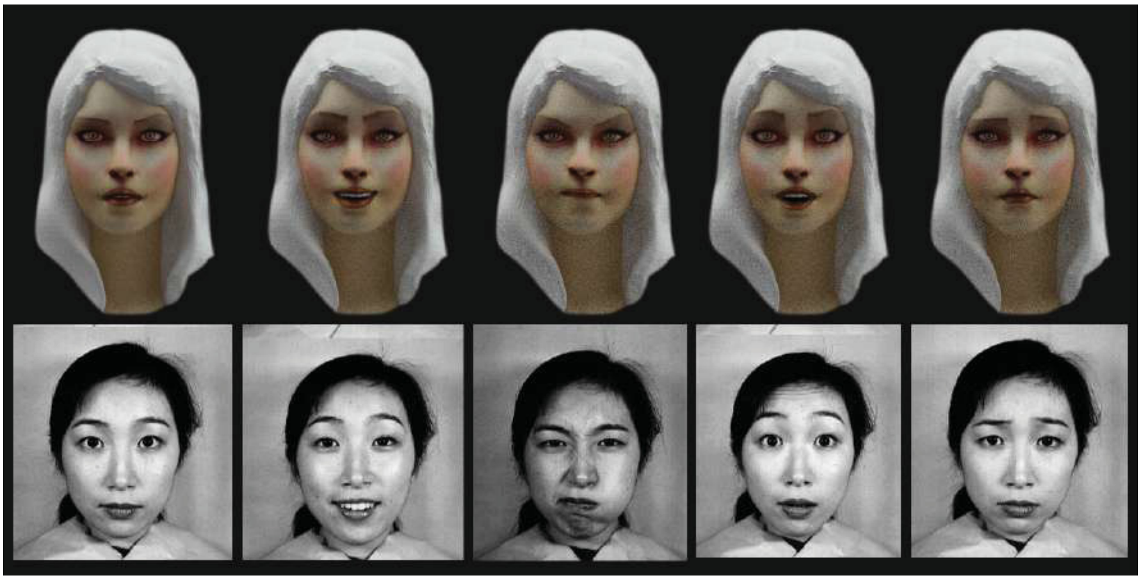


Figure 3.1: Basic emotions represented in the virtual environment. Top row shows faces built with the 3D mesh, from left to right: Neutral, Joy, Anger, Surprise and Sadness. The bottom row shows the samples of the dataset (Lyons et al., 1998) used as a reference for deformations in the 3D mesh.

animation context a clean 3D model is fundamental so that an optimized facial model can be applied to ease the manipulation of the parameters and values of each landmark, or associated facial point based on the points of the MPEG-4 FP model.

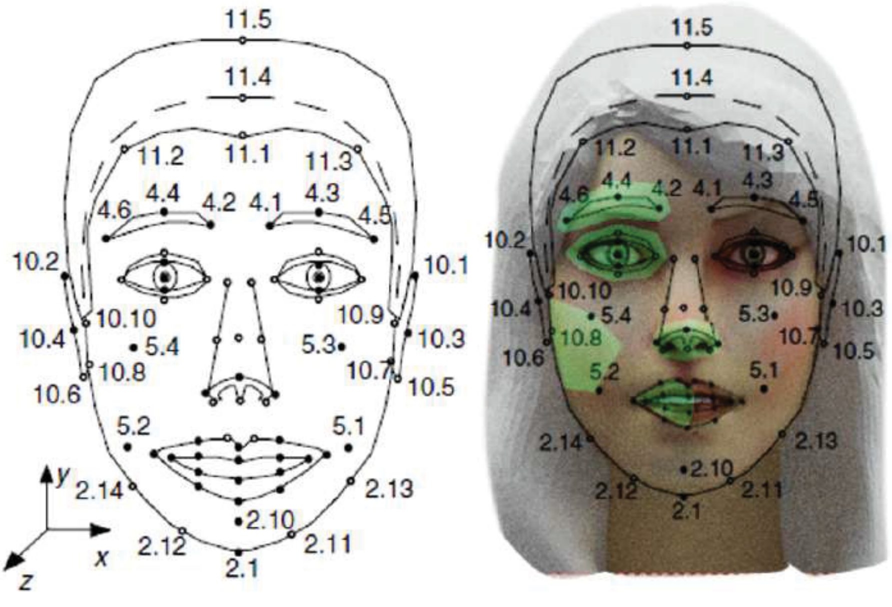


Figure 3.2: Facial Regions, in green, and facial action parameters based on MPEG-4 model. The defined regions include key areas in face on expressions synthesis process, that are: Forehead, Mouth, Cheek, Eyes, and Nose. Geometric symmetry is assumed aiming to optimize the experiments.

The facial regions allow the behavioral tests of the landmarks to be observed in isolation, identifying areas of the face that are more or less relevant in the process of representing the base expressions. Moreover, the regions Obaid et al. (2010) allow the parameters to be independently evaluated associated only by the hierarchy defined for the synthesized signal, approaching the facial model defined in this Chapter with the already used models for manual signal representation.

The deformations were built using blend shapes (Dailey et al., 2010) based on the Japanese Female Facial Expression (JAFFE) dataset, which provides an archive with samples of the basic emotions interpreted by 10 Japanese subjects (Lyons et al., 1998). This dataset classifies the images using semantic ratings averaged, which defines expressions by points of interest values, statistically identifying which emotions each image is and its intensity generally (the displacement of the facial muscles and facial points of interest define the expression represented and the value of intensity). JAFFE dataset includes sixty subjects and uses a five level scale classifying each image which a value form the base expressions Happy, Sad, Surprise, Anger, Disgusting, and Fear. According to authors, Fear in the only base expression which can not be accurately evaluated in a controlled environment, pointing more subjective values for this expression.

In order to compare facial interest points and reinforce values and interest points from the JAFFE dataset, the Averaged Karolinska Directed Emotional Faces (AKDEF) were used too. This dataset uses parameters similar to the JAFEE dataset where values are associated in each image based on the distances defined in each facial expression based on interest points. The AKDEF provides seventy images of human facial expressions. The set of pictures contains an averaged female and male viewed from five different angles displaying seven different emotional expressions, the same basic emotions for Ekman model (Lundqvist, 1998). Figure 3.3 shows an example of the AKDEF base where the representation of base expressions can be noted with examples of expressions represented in the datasets with specific values for synthesized expression by percentage value.

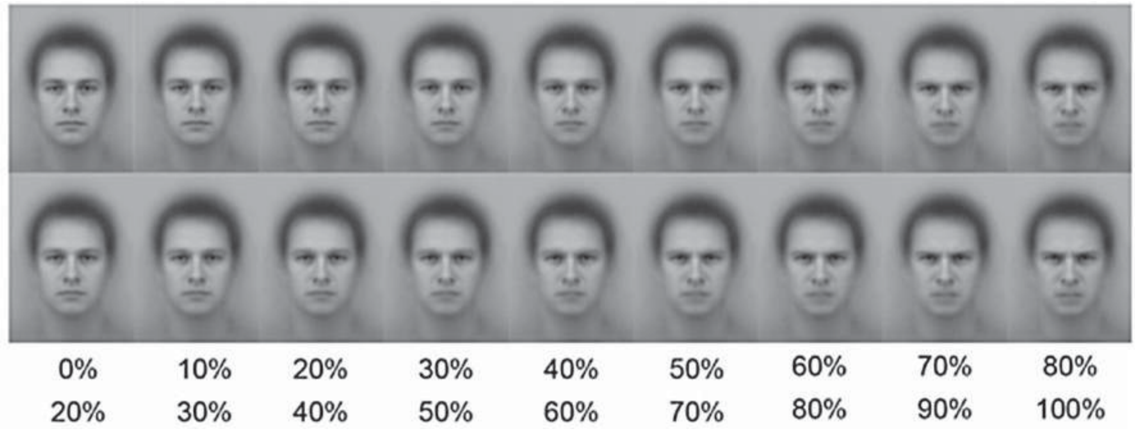


Figure 3.3: Lundqvist (1998) AKDEF dataset sample. The images sequences follows the percentage presented below as intensity in the representation of the expression, based on the position of the points of interest in the human face. The points of interest are the facial muscles that when displaced indicate the expression or emotion that the face intends to express. The percentage values represent the intensity in which the expression is represented in each image, 0% for the neutral expression and 100% for the angry base expression.

The JAFEE dataset uses a similar structure to the Table 3.1 where the values for each base expression referenced in the input image calculated based on the neutral facial model and its specific displacements. Table 3.1 includes two examples: the Figure 3.1 in dataset, of code *KM-NE1*, has a value of 2.87, on a scale of 1 to 5, in the classification of joy expression and 2.10 of surprise while the code image *KM-NE2* has a lower rating value for the surprise expression.

Table 3.1: Example of values applied in the JAFEE dataset. The first row, composed by the the base expressions Happy (HAP), Sadness (SAD), Surprise (SUR), Anger (ANG), Disguise (DIS), Fear (FEA), and the name of picture (PIC), while rows 2 and 3 present the values classified for each input image. The values represent the intensity of the expression varying from 0 to 5 for less and more expressive, being calculated by similar criteria to the AKDEF dataset where the position of facial regions define the intention in the representation of the emotion.

	HAP	SAD	SUR	ANG	DIS	FEA	PIC
1	2.87	2.52	2.10	1.97	1.97	2.06	KM-NE1
2	2.87	2.42	1.58	1.84	1.77	1.77	KM-NE2

The AKDEF base, otherwise, uses a structure as described in Table 3.2, defining in addition to the base expressions, the angle of the registered face. This dataset offer samples for both sexes. The values that define the classified expression consider spatial displacements of facial points of interest, based on the values extracted from the neutral expression.

Table 3.2: AKDEF dataset example: the sex of the subject, the expression class and the position of the face considering the camera direction. As in the JAFFE base, each image can have semantic tags associated with more than one expression considering interpolation between classes.

AKDEF Sample: FAFFL.JPG		
Gender	Expression	Angle
F = female	AF = afraid	FL = full left profile
M = male	AN = angry	HL = half left profile
	DI = disgusted	S = straight
	HA = happy	HR = half right profile
	NE = neutral	FR = full right profile
	SA = sad	
	SU = surprised	

The use of the blend-shape technique, that consists of moving the geometric mesh based on a reference, follows the recommendations of Cetinaslan e Orvalho (2018), applying the deformations in precise points in order to faithfully follow the facial datasets JAFFE and AKDEF. The formal models linked to landmarks, defined by the vertexes representing the points of MPEG-4 model, contain parameters for location, movement, contact, and other characteristics that represent a particular pose or configuration of a body part (Punchimudiyanse e Meegama, 2015). In the case of facial expressions, we can define the activation of an emotion or an interpolation between more than one expression. This is usually done using keys to body parts and values for the movement of the action. In the case of face and hands, a more specific set of parameters can be specified assuming controllers for each region of the face as well as values indicating new coordinates for their vertexes.

Tracking points were created in the neutral expression images using the defined facial parameters and, based on their values, blend shapes were animated relating the displacement of landmarks with the points of the geometric mesh. Each sequence representation of facial expressions was built which around 40 frames, following Xue et al. (2015), where they assumed that audiovisual model cadence uses 25 fps (frames per second) in PAL-M system, 30 fps in NTSC model and 60 fps in digital format.

The coordinates for each expression points of interest in face region were defined based on the values found in the two bases. Following the semantic values identified in each base

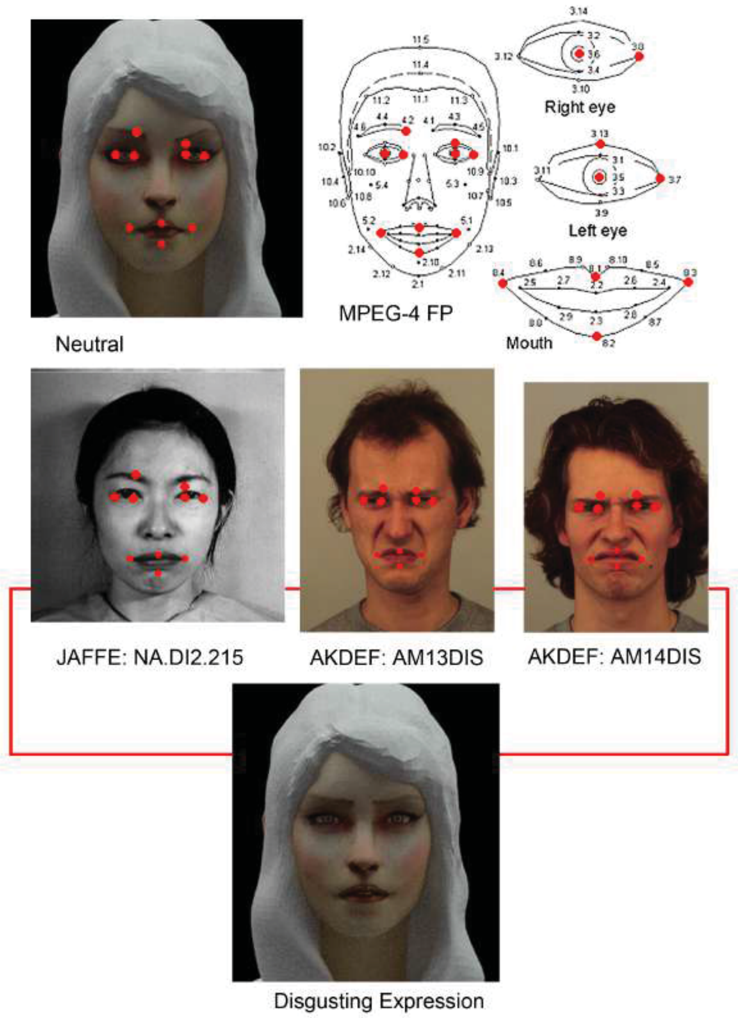


Figure 3.4: Generation of the synthesized expression based on the subjects of the datasets used AKDEF and JAFFE. The red dots were defined based on the MPEG-4 and applied to the 3D avatar, and based on the position of the points traced in the input images, the blend shapes were constructed as animation for the base expressions. In the image, inputs like the three figures displayed (NA.DI2.215, nomenclature used in the JAFFE dataset that defines the subject (NA), expression (DI2) and image number (215), and the AKDEF dataset examples that represents the subject gender (AM), identification (13 and 14) and expression (DIS)) were used to define the displacements of each point for the generation of the disgusting expression in the 3D avatar.

expression, modifiers were applied to the 3D geometric mesh by deforming the model in order to fit in the average values as shown in Figure 3.4. With the blend shapes of the applied base expressions, coordinates of specific points on the face of the 3D avatar, linked to controllers, can be identified as well as predictions of positions for interpolation of expressions.

The next experiments were developed in order to better understand the behavior of the landmarks of the face in the process of synthesis of base expressions.

3.1.0.0 Neutral Expression Deviation Factor (NEDeF)

In this section, an analysis of the displacement of the landmarks, associated with the regions deformation aiming to identify the most affected regions on the 3D mesh for each base expression.

A novel metric is proposed to measure the global normalized region deformation, representing the influence on the mesh for each expression, defined by the Neutral Expression Deviation Factor (NEDeF) ((Eq.3.1)):

$$\frac{\sum_{v_{ir}\in R_r}\frac{|d_n v_{ir} - d_e v_{ir}|}{d_n v_{ir}}}{NV_r} \qquad \text{(Eq.3.1)}$$

Where, for each expression e a measure of the distortion relative to the neutral expression n is computed. This is performed, for each facial region R_r, by the normalized sum of differences of the Euclidean distances in the 3D space from the centroid region to each respective landmark. A second normalization is computed considering the number of landmarks (reference points) defined for each region NV_r. The distances were taken in absolute values because the distortion of the regions is assumed to be additive.

The spatial location of the landmarks in the same region define respective centroid. The absolute values of the displacements were used to calculate the NEDeF ((Eq.3.1)) considering that one distortion in the mesh should be considered in any direction in the virtual environment in order to evaluate deviation from the neutral expression. Were considered the extracted coordinates of landmarks added to the shape keys values that control the blocks of the intensity of expression modeled in the 3D object, by the coordinates of its vertices or controllers position (3.5).

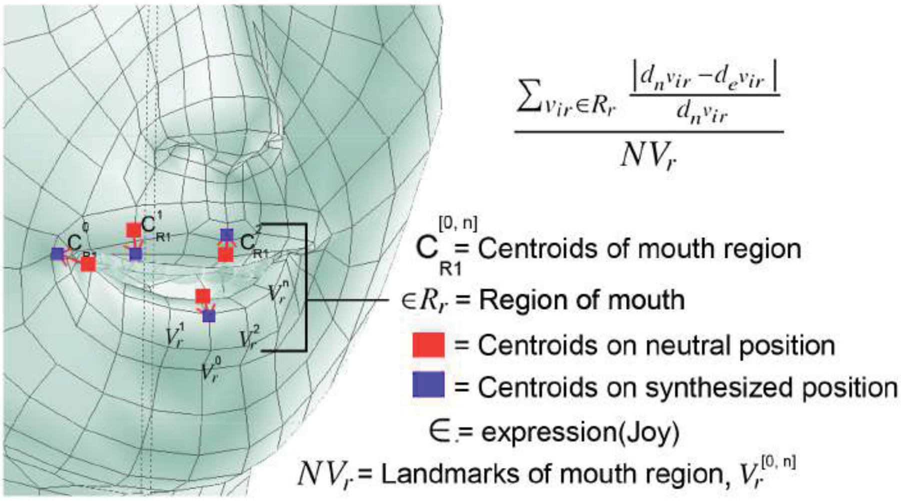

$\frac{\sum_{v_{ir}\in R_r}\frac{|d_n v_{ir} - d_e v_{ir}|}{d_n v_{ir}}}{NV_r}$
C^0_{R1}
C^1_{R1}
C^2_{R1}
V_r^n
V_r^1
V_r^2
V_r^0
$C^{[0,\,n]}_{R1}$ = Centroids of mouth region
$\in R_r$ = Region of mouth
🟥 = Centroids on neutral position
🟦 = Centroids on synthesized position
$\in$ = expression(Joy)
NV_r = Landmarks of mouth region, $V_r^{[0,\,n]}$

Figure 3.5: NEDeF parameters example applied in 3D avatar used in this work.

Table 3.3 shows the NEDeFs of the regions and their distortion points compared to the same landmarks with the synthesized expressions, together with the normalized values of intensity of influence in the 3D mesh were extracted in each region.

Table 3.3: NEDeF of the basic expressions for key regions, indicating the influence of each region in the construction or animation of each expression.

Geometrical Comparison of Facial Expression Regions					
	Forehead	Eyes	Cheeks	Mouth	Nose
Joy	0.118	1.000	0.800	0.791	0.376
Anger	0.628	0.900	0.953	0.702	0.080
Surprise	0.363	0.460	0.000	0.970	0.000
Fear	0.250	0.484	0.673	1.000	0.095
Sadness	1.000	0.980	0.307	0.400	0.091
Disgust	0.730	0.050	0.667	1.000	0.000

The results shown in Table 3.3 allow the identification of the more affected regions for each emotion. The forehead region is a highlight in the expressions of anger and sadness and the mouth region is of great importance in joy, surprise, and fear expressions. For the joy emotion, the regions with more distortion in the mesh were the cheeks, on the negative expressions (anger, fear, sadness) nose region tends to have a more noticeable change in comparison with the positive emotions (joy and surprise). It is important to mention that, based on Plutchick's model, emotions can be interpolated (Szwoch, 2015), thus, all expressions in the synthesis process can be modulated by an intensity factor.

3.1.0.0 Facial Parameters Principal Components Analysis.

The next facial experiment aimed to identify the principal components of the face on emotion synthesis process. For this, five samples were used relating facial regions. The landmarks of the face represent the variables with their Euclidean values ranging from a neutral expression to the extracted synthesized expressions of calculations of their coordinates based on the distance from its centroid directed to the value of blend shape. The co-variance matrix was calculated as the eigenvalues and eigenvectors of the average vector of the samples. The principal components analysis algorithm reduce the dimensions of a large amount of data pointing out the variance between correlated variables, in this case, the facial landmarks.

Besides the fact that the landmarks displacement represents the influence on the geometric mesh, knowing the correlation between the landmarks by Euclidean distance can help to define what facial region is most relevant in the synthesis process for each emotion. For this PCA algorithm can be used in order to calculate the explained variance ratio of the landmarks given by the EVR value of the principal components. Then the data was rearranged in a Hotelling matrix, used to solve such multivariate testing problems, in order to obtain their principal components. Figure3.6 depicts a graph representing the landmarks in two-dimensional space and their representation after applying the PCA algorithm.

The Explained Variance Ratio (EVR) was obtained based on the eigenvectors and eigenvalues, where it was observed that 58 % of the variance of the data is in the direction of the principal components of the Forehead region, being the Mouth region with the second most expressive value with 26 %, followed by Cheek's with 11 % of the variance of the data directed to its components. The fact that the forehead region has a higher value of variance indicates that the most identifiable displacements in the emotions may occur in this region, which together with the region of the mouth may be sufficient to represent an identifiable emotion.

The Eye region had less expression in the tests, followed by the nose region with had the EVR value lowest than 1 %. The EVR values indicate in which regions the simplification can be more effective, besides pointing out, in conjunction with the NEDeF result values, the relation between a facial region and a synthesized base expression. Also, these results enable a landmark behavior discussion, giving values for each trajectory in geometric space.

In order to support the previous experiments, Factor Analysis Algorithm (FA) was applied. This experiment allows the understanding of the facial landmarks behavior. The same PCA variables of the previous analysis were used as facial regions and facial expressions, observing the weights of their relations. Factor Analysis algorithm allows to identify the main factors among correlated data, in this case to indicate which are the most important facial regions in the synthesis of the six base expressions, also calculating the variance between the data considering its displacement.

Figure 3.7 shows the data distribution where the variables are represented by the five facial regions of the previously defined model, and the initial factors as the expressions for basic emotions (Joy, Anger, Sadness, Fear, Surprise, and Disgust). The values distributed in the data matrix were obtained from the relative displacements from the neutral position to locations after the expression synthesis. The relative displacements were calculated taking as reference the landmark's centroid for each region.

The objective of the FA application was to find the co-variance between the regions in the execution of an expression using the weight of the relation between the data of the geometric mesh, defining the most relevant regions in synthesis of each base expression. The simplification of extracted factors should be useful in identifying which regions are most expressive in the observed synthesis process, ignoring the less influential region in the synthesis process or interpolating to a less expensive vertex offset, manipulating fewer centroids by synthesis.

This process, as well as the PCA algorithm, consider the blend-shapes modeled based on the synthesized outputs in 3D avatar for each base expression, using average parameters of the JAFEE and AKDEF bases, and may vary for other facial bases.

In order to confront the PCA results, the Factor Analysis algorithm was implemented and applied to the facial landmarks for each expression. The normalized values for the analyzed factors eigenvalues are 0.79 for Forehead, 0.0 for the Eyes, 0.85 for Cheeks, 1.0 for Mouth, 0.13 for Nose region, which shows that using the 5 facial regions and the six base expressions of the Ekman's model, we can consider three main factors. The graph in Figure 3.8 shows the reduced

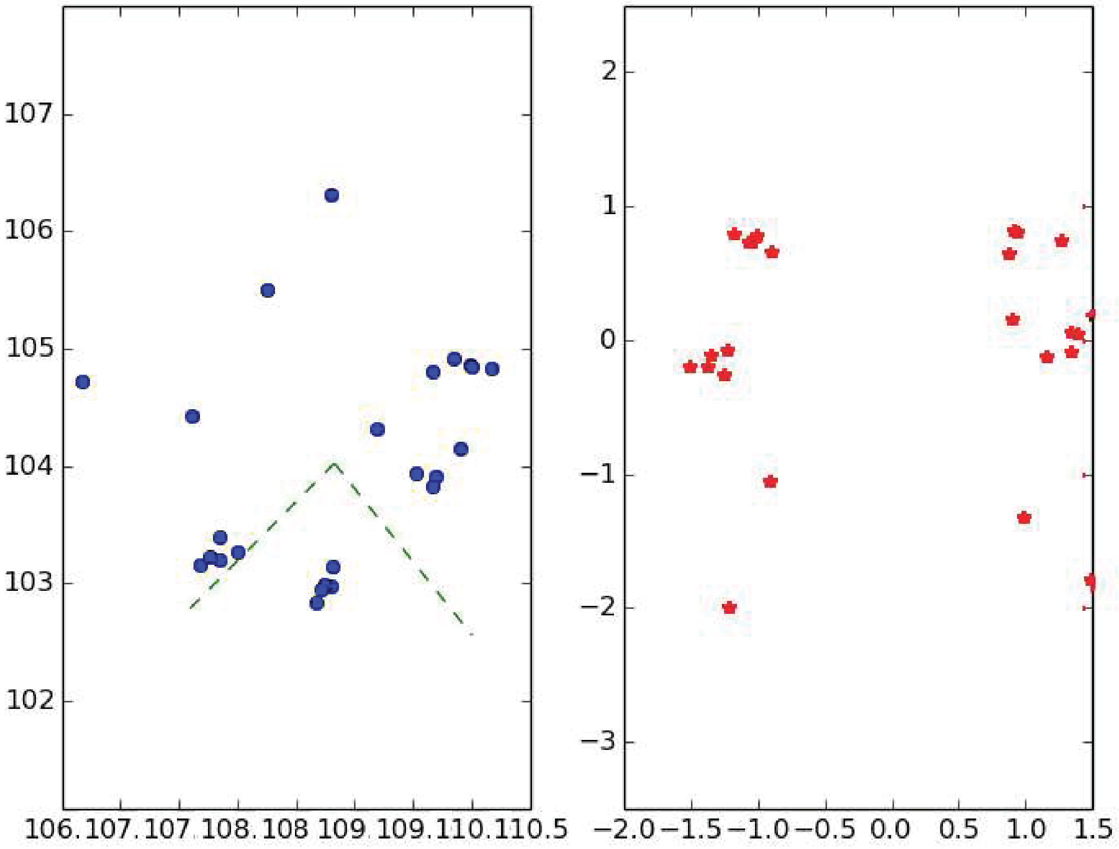


Figure 3.6: PCA applied to 3D mesh landmarks vector: It's subtracted the mean of facial landmarks values. According to EVR components of Forehead and Mouth regions had most expressive displacement values represented by the amount of points following the dotted lines. These lines at the left image show the maximum variance direction of the first and second principal components, as Forehead and Mouth landmarks, and the right image shows the points represented in the new principal component's base, aligned with reduced dimension based on the landmarks of the two main regions.

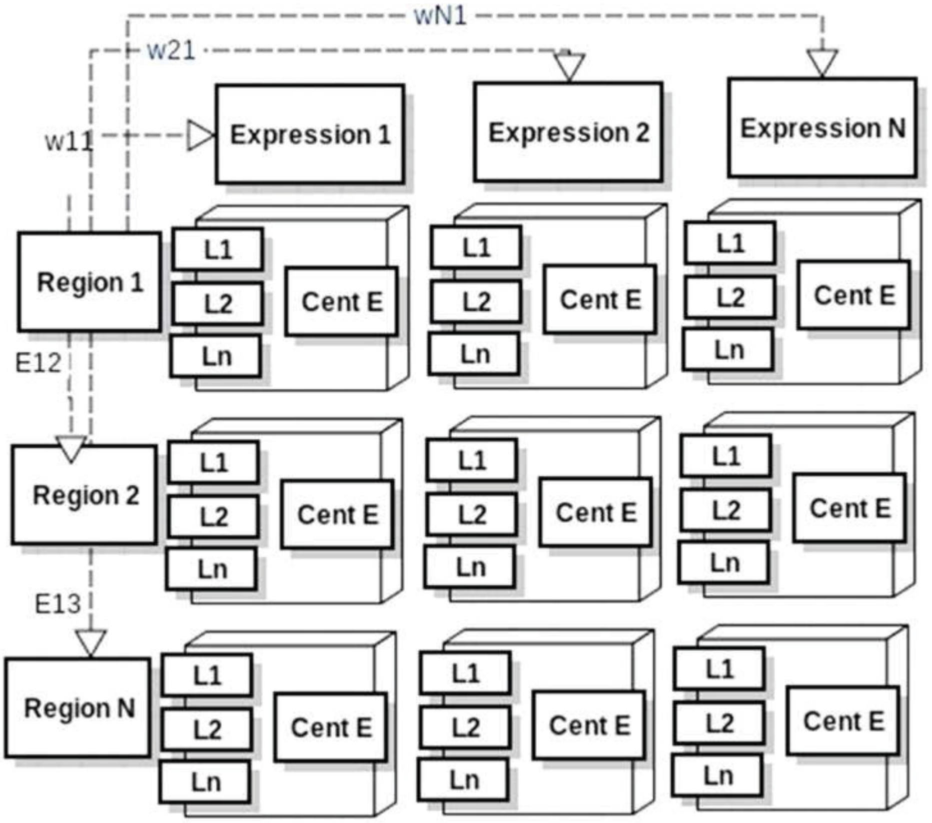


Figure 3.7: The data model for FA Algorithm, where centroid coordinate absolute values are calculated between facial region and expressions based on their weights (parameter *w* based on NEDeF influence). In the image, the parameters *L1* to *Ln* represent the landmarks and *Cent E* refers to the centroids of the expression.

factors based on the eigenvalues, where the absolute values of regions of mouth and nose can be reduced in a single factor, as well as values cheeks and forehead.

This means that, in the synthesis of base emotion expressions, using the landmarks based on the MPEG4-FP model and the regions defined in the previous sub-section, we can point that the Forehead and Cheeks regions had a similar displacement expressiveness in the geometric mesh. The Mouth region was the most expressive with the most noticeable details of geometric change and Eyes and Nose can be considered as a less influent factor with less perceptive displacement.

Considering that the landmarks of the mouth have a more remarkable displacement especially due to the expressions of fear and surprise, the results point to the cheeks as the second region with the greatest relation between the displacement of the facial landmarks in the synthesis of the expressions, which is very relevant since that most part of the systems that use 3D avatars, mentioned in previous chapters, when mapping the face, generally consider animations for the mouth and eyes and completely ignore the region of the cheeks. This experiment points out a relevance to the vertices associated with this region once they have an important influence on the main positive and negative expressions, which are joy and anger.

The results of the PCA and FA tests indicate that the landmarks can have their dimensions reduced based on the analysis of displacement in Euclidean space and corresponding expressiveness in the synthesis of emotions. Trajectory extraction methods can be explored based on which facial regions have more significant displacements in expressions animation. This analysis can be extended using methods such as Euclidean Distance Variance Matrix Analysis, or

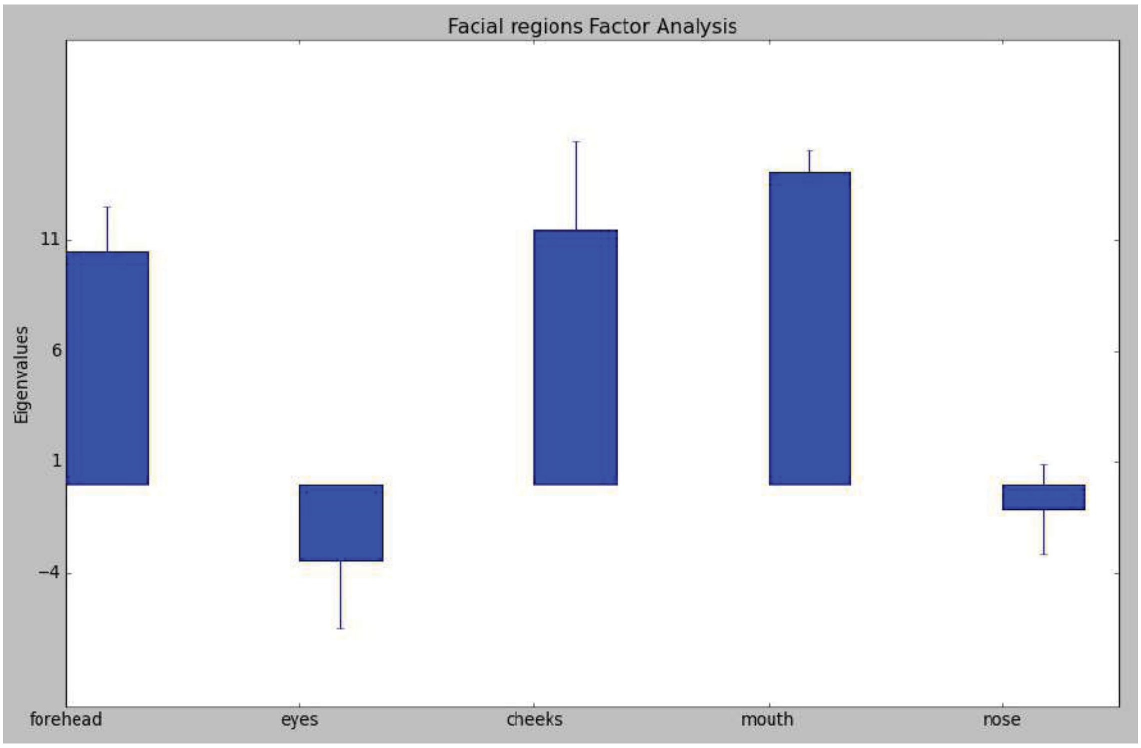


Figure 3.8: Main factors for facial regions. The five values, from left to right, represents the absolute values for regions: Forehead, Eyes, Cheeks, Mouth and Nose

Steepest Descent and Cyclic Coordinates through the behavior of the centroid trajectory or ratio of the largest eigenvalue to the smallest eigenvalue of the Hessian matrix (Hovden e Ling, 2003).

Other elements that could be considered are the use of normal textures aiming facial animation optimization in virtual environment (Zhao e Zhang, 2008). Facial expressions in sign language use facial details to define important information that can change the meaning, for example, doubt for irony expression. Small deformations in the geometric mesh can be replaced by changes in texture using Normal Bump or Displacement channels to represent that information (Yao et al., 2018). In that case, the changes in the model are simulated by refractive of light changing diffuse layer coordinates of a texture. These techniques, however, aren't used here because it is possible to generate the interpolations without depending on a parameter based on textures which is out of the scope of this work.

3.1.1 Concluding Remarks

Facial expressions have a fundamental importance in the transmission of a sign language message. In this chapter a novel BSL Facial Model that comprises the main facial regions is presented, expanding the MPEG-4 FP model and the concepts presented by Kacorri e Huenerfauth (2014), Lemaire et al. (2011), Obaid et al. (2010) and Lv et al. (2015) where they were defined centroids and controllers for the main points for deformation of the avatar 3D mesh and define the base expressions, according to the semantic values presented by Lyons et al. (1998), Lundqvist (1998), and Huenerfauth et al. (2011). Also, the NEDeF model was proposed to identify the influence of each region on the synthesis of the base expressions.

The base emotions are generic and do not need a re-mapping of points and synthesis in the future. The only adjustment in this step will be for controllers when using another 3D model. The next objective in order to define more precisely the facial landmarks behavior is to analyze the interpolation process of emotions, as well as their coordinates in a spatio-temporal model.

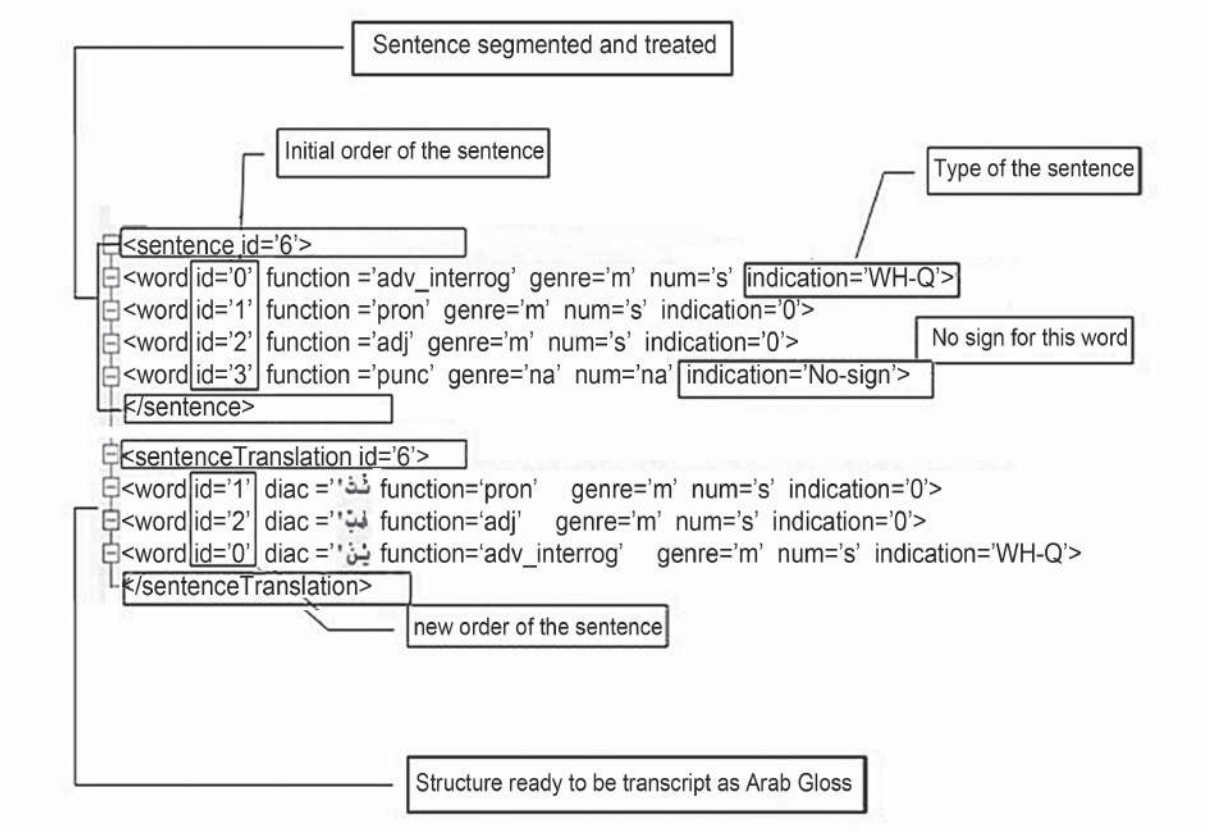
With the models and methods proposed and constructed, the Chapter 4 presents the data parser model proposed for the integration of facial expressions into formal models of sign languages used for manual parameters and Chapter 5 reach a solution where the BSL Facial Model is adapted to 4D trajectories, based on a modeling and experiments that define the behavior of the facial landmarks.

The NEDeF concept is extended in Chapter 5 in order to optimize the representation of facial expressions for sign languages, defining the behavior of facial regions and supporting a method for expression interpolation generation.

4 DATA EXTRACTION AND PARSER MODEL

The following sections present the experiments for data output and parser process. Through a sequential geometric landmark tracked, a hierarchical data export method was proposed, working as a computational formal model for facial landmarks easing the increment of new parameters or the interpolation of expressions by changing parameter values defined for base emotions. The data parser is fundamental to define a representation of the face parameters in a similar format as used for manual sign parameters in sign languages systems, and thus to be able to extend the representations of emotions in a defined and controlled way.

The follow experiments use generic parameters of a 3D facial mesh, that can be adapted in existing systems that have mapped signal languages parameters in computational description such as the CORE-SL model shown by Iatskiu et al. in 2017, where the manual parameters already have a defined structure that can be increased with the facial models shown in this work. Another computational representation of sign languages is the formal model presented by Aouiti et al. in 2017 that uses a hierarchical configuration to represent, for example, the sentence "Where are you going" adding value for signal parameters for each word in Arabic Sign Language, as seen in Figure 4.1.


Sentence segmented and treated
Initial order of the sentence
Type of the sentence
No sign for this word
new order of the sentence
Structure ready to be transcript as Arab Gloss

```
<sentence id='6'>
<word id='0' function ='adv_interrog' genre='m' num='s' indication='WH-Q'>
<word id='1' function ='pron' genre='m' num='s' indication='0'>
<word id='2' function ='adj' genre='m' num='s' indication='0'>
<word id='3' function ='punc' genre='na' num='na' indication='No-sign'>
</sentence>
<sentenceTranslation id='6'>
<word id='1' diac ='' نْتَ function='pron' genre='m' num='s' indication='0'>
<word id='2' diac ='' هِبْ function='adj' genre='m' num='s' indication='0'>
<word id='0' diac ='' يْنَ function='adv_interrog' genre='m' num='s' indication='WH-Q'>
</sentenceTranslation>
```

Figure 4.1: Example of XML parameters structure for Arabic Sign Language sentence (Aouiti et al., 2017). Each word receives intensity values and represents independent parameters, organized in hierarchical dependence, relating the values of its parameters in order to represent a sentence.

The focus of this Chapter, therefore, are not models for the facial data or the generated trajectories of the expressions, but the system of extraction and parser of the geometric data.

4.1 3D MESH COORDINATE EXTRACTION BASED ON TEMPORAL FEATURES

Extract blend shapes information of a 3D human face consists of a vertexes coordinates sequence that depends on factors such as the process of synthesis, that define the translation and rotation of the geometric landmark and key-frames, that define the changes based on an observed animation.

The next subsections present methods for dynamic object recognition in virtual environments in order to generate coordinates that can be written to an external file allowing the implementation of automatic synthesis exportation and parser system for sign language parameters. The objective of those experiments is to propose a model of data extraction that can incorporate the NEDeF model and spatio-temporal analyzes that will be presented in the next Chapters. The XML model presented can also be adapted to more specific parameters.

4.1.0.0 Landmark tracking in 3D facial mesh

Two inputs are commonly used for facial feature definition: a capture and classification of sequential facial images from video inputs, or, animation in a virtual environment of an avatar with human characteristics. Those patterns are usually based on a formal model that defines the facial parameters (Kacorri et al., 2015) (Yan et al., 2008).

These parameters can be tracked during the sequence of inputs, constructed using techniques such as blend shape for virtual models (Casas et al., 2016) (Sagar, 2016), that link controller in the virtual environment to points of interest defined in the inputs. The input that contains facial characteristics are synchronized to landmarks, in that case, geometric vertexes or controllers (virtual objects that control the displacement of one or more vertexes).

The first step to build feature extraction algorithms is to identify what information is expected to obtain. The sequential information is extracted from readings of landmark position in time into a known space (Sun e Yin, 2008), in this research, facial landmarks, and their displacement coordinates. In 3D objects, any surface or feature analysis can be made obtaining information from their geometric mesh and their coordinates. In cases where the environment is relevant, the normal information of the faces can be considered which contains useful data about the refraction of light that may help clarify the position of the point on the environment around them.

Since the objective at this time was only the creation of the coordinate data extraction and parser system, a sequence of head movements was created using landmarks in the face on specific points that are the eyebrows, mouth corners, a few points on the contour of the face, forehead, and nose, based on the model used byCao et al. (2014). These parameters were applied to a video input in order to generate a simple trajectory without considering complex blend-shapes or facial expressions. The intent of this input was just getting the landmark position of a dynamic object and the temporal changes such as the experiments performed by Wu in 2016. Landmark tracking was pointed using Harris Corner algorithm and a Threshold filter, suited to a few points and compared to each image sequence position by pixel gray-scale (Figure 4.2).

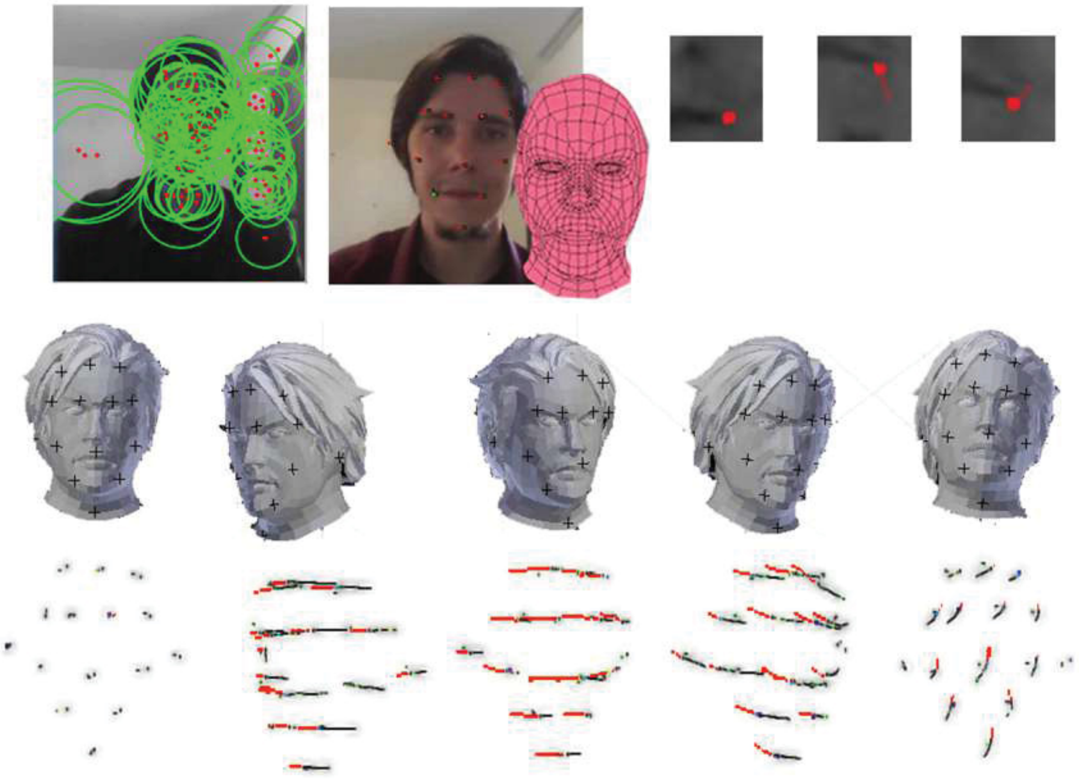


Figure 4.2: Dynamics landmark acquisition: In a cloud of points was defined main facial points of interest based on MPEG-4 model and Cao et al. work. For this experiment were built a tracking for individual points, followed by synchronization with the low poly 3D model. The temporal path was identified comparing the gray level associated with geometric controllers and linked to vertexes in avatar face. The bottom sequence shows the process of tracking the facial landmarks and the temporal readings of the landmarks extracted of the geometric mesh.

With this preliminary tracked data, a data parser system was developed, presented in the next subsection, where the export and storage process for an external file is explored.

4.1.0.0 Geometric Mesh Data Extract

With the facial points coordinates, a data extraction and parser system was constructed using the Irrlicht Graphic Engine and external files read and write functions in order to generate an XML output in the hierarchical format used in formal computational models of sign languages. All those methods were implemented in C ++ programming language.

Based on the movements tracked in the input video, the coordinates of the 3D mesh points of interest are ordered in a structure that follows the signal pattern with the hierarchy based on expression, facial region, and then the controller point and its Euclidean coordinate values in the interval $[a, b]$ that gives the position on the animation key-frames.

In the case of facial expressions, an emotion its spatial coordinates and temporal information (based in the animation key-frames and geometric blend-shapes), as follows the XML slice below, where it is possible to observe what would be an example of expression Joy that receives parameters such as the category (positive or negative), position of frames in the animation and coordinates for the centroids of each region:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified"
   elementFormDefault="qualified"
   xmlns:xs="http://www.w3.org/2001/XMLSchema">
   <xs:Expression = "Name">
      <xs:sentence: "Joy", srclang="br" lang="bsl"
         srcsentence="Positive Expression", "Neutral"">
      <xs:sequence>
         <xs:key-frame> value: 25 </xs:key-frame>
         <xs:region> value: 1, "mouth"
            <xs:COO>
               [:] List of coordinates for Centroid Euclidean Distance.
            </xs:COO>
         </xs:region>
         <xs:region> value: 2, "cheeks"
            <xs:COO>
               [:] List of coordinates for Centroid Euclidean Distance.
            </xs:COO>
         </xs:region>
      </xs:sequence>
      </xs:sentence>
   </xs:Expression>
</xs:schema>
```

The parameters will be defined according to the spatio-temporal data model presented in the Chapter 5. This XML model can be adapted to existing computational formal models for Sign Languages as well as enhanced with new facial information and finer spatial controllers. This model can support the 4D regions outputs based on the centroids trajectories of each region.

A parser method in the graphics engine was developed aiming to recognize the elements locally. So, they can be worked in sequence with Sign Language current systems to generate the process of 3D avatar facial expression animation. The parser method was implemented and executed using the Irrlicht Graphic Engine and OpenMP parallel programming methods. The complete process can be paralleled since the method are divided into independent similar steps, and the amount of information handled tends to be large. The Figures 4.3 and 4.4 show the process fluxogram.

The idea of the data output process presented in Figure 4.3 works in a similar way that the automatic signal synthesis systems presented in Chapter 2 (Othman e Jemni, 2017) (Kaur e Singh, 2015) (Ahire et al., 2015) (Kacorri, 2015) and can be performed using parallel functions for data classification. The classification aims to identify the structure of the coordinates found, by facial regions, or facial expressions extracted.

In Figure 4.4 the data referring to landmarks, or their coordinates in a virtual environment, as well as auxiliary information such as their facial region, are passed as parameters for a function that validates its structure, writing the information in the external file. The idea of the external file is to follow a simplified textual format where the parameters can be changed depending on the new synthesis defined by the formal model.

The parser method can be used in applications that already synthesize animations of sign language signals integrating facial parameters to their avatars, expanding context information and emotions transmitted messages. The main idea of this system is the incentive to include these syntheses since parameterized models follow a similar data structure for manual gestures.

4.1.1 Discussion

A dynamic facial landmarks extraction based on the normalized Euclidean coordinates values may provide temporal moves information. These methods will be used in the next Chapter for a

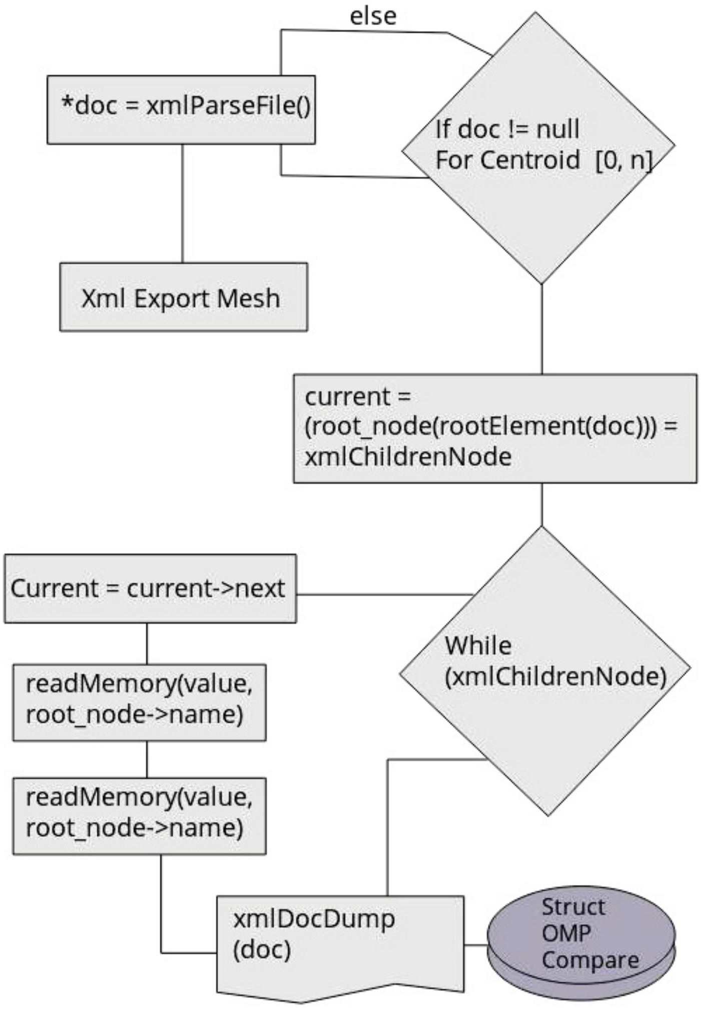


Figure 4.3: Parser for mesh data process. The XMl data are read and stored in the internal memory and an process to compare and classify geometric data are executed. The parser system may consider the signal hierarchy by comparing the coordinates of each facial region in parallel in order to identify, by the read values, the signal or parameter to be described. There is no standard for this model, but it can be easily adapted to formal models already used in the literature in order to integrate the structure and parameterization used in the current systems.

facial expression interplation and simplification methods. The parser methods presented in this Chapter can extract data in a simplified and functional model and enable current formal models (Iatskiu et al., 2017) for Sign Language Systems to integrate the facial parameters proposed along this research.

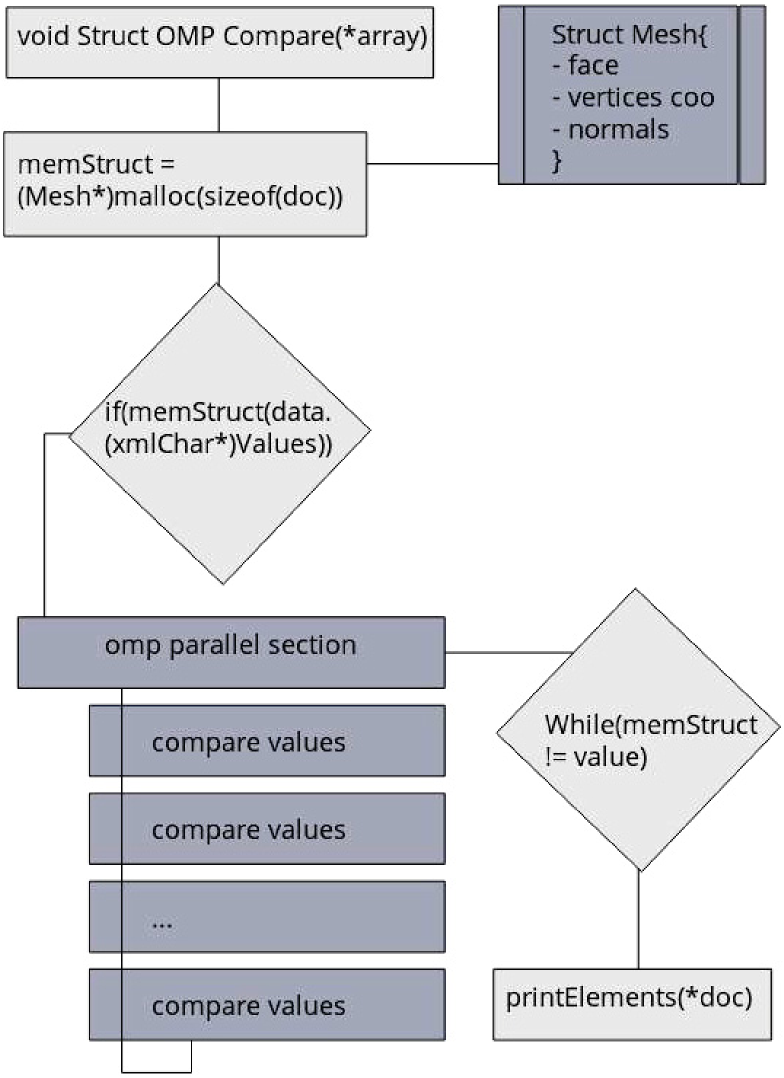


Figure 4.4: Sequential 3D mesh Struct Comparison. Once elements are in memory the values of their coordinates are evaluated sequentially. If an automatic signal generator system uses external XML files with hierarchical parameters, each face region could be read and its values compared with a signal definition simultaneously to generate facial animation.

5 MODEL FOR FACIAL LANDMARKS TRAJECTORY BEHAVIOR

This chapter proposes a Spatio-Temporal representation for facial expression generation, based on NEDeF model. With the methods presented in the following sections, it is possible to represent the facial landmarks trajectory along the time when facial expressions are produced. Section 5.1 presents models used to represent temporal objects, followed by section 5.2, where a novel model is proposed for representing the Brazilian Sign Language (BSL), data in 4-dimensional (4D) space.

Landmarks coordinates and their respective trajectory extraction process are presented in the sequence, supporting the synthesis of expressions for a 3D avatar dedicated to Sign Language communication. Finally, the third section presents a validation of the proposed model based on human evaluation of generated expressions.

5.1 SPATIO-TEMPORAL DATA MODEL FOR FACIAL LANDMARKS

Considering the models presented in Chapter 2 and related spatio-temporal data (Sheidin et al., 2017), (Xu et al., 2017), (Hongjian et al., 2016) and (Aldrich, 1998), this section defines a practical model for the 4D Facial Landmarks representation. This model is applied in the trajectory extraction of the facial regions in order to identify the behavior of the geometry mesh of the avatar for emotion synthesis aiming optimized expression generation.

The MPEG-4 + Facial Regions model, based on the facial landmarks shown in Chapter 3, can be extended with the concepts concerning spatio-temporal data. The centroids pointed for the five facial regions (nose, mouth, two eyes, cheeks) are dynamic geometric points that generate a sequential trajectory curve in the process of synthesizing expressions through the 3D avatar. This framework, constituted by landmarks associated with facial regions,and their spatio-temporal trajectories, allows to observe them as 4D landmarks, considering their trajectory in a temporal interval in a controlled space. This section defines a model for the facial landmarks used to extract trajectories based on the fitting to a curve model increased by the values pointed by the PCA analysis and centroids trajectories as shown in Figure 5.1.

Based on Bezier parametric curve (Dias et al., 2016) (Aldrich, 1998), formed by a B-Spline or a junction of several tabulated values of polynomial functions, for degree k with $k - 1$ control points, the model follows the (Eq.5.1):

$$B(t) = \sum_{i=0}^{n} N_{i,j}(t) P_i \tag{Eq.5.1}$$

Where in a uniform B-spline $B(t)$ using a sum Bernstein Polynomials $N_{i,k}(t)$ as a basis function and P_i as a control points $(P_0, P_1, ..P_n)$, in this work centroids of facial regions. The basis function is defined using Cox-de Boor recursive formula considering t as index of temporal points in pairs sequence (Aldrich, 1998) ((Eq.5.2)):

$$N_{i,j}(t) = \frac{t - t_i}{t_{i+j} - t_i} N_{i,j-1}(t) + \frac{t_{i+j+1} - t}{t_{i+j+1} - t_{i+1}} N_{i+1,j-1}(t) \tag{Eq.5.2}$$

The values of the points represented by t is a part of the sequence called Knot Vector and determine the basis function that influences the shape of the B-Spline trajectory. Knot vector is represented by $t = (t_0, t_1, ...t_n)$ in range $t \in [t_0, t_n]$ (Aldrich, 1998). Each centroid is a spatial

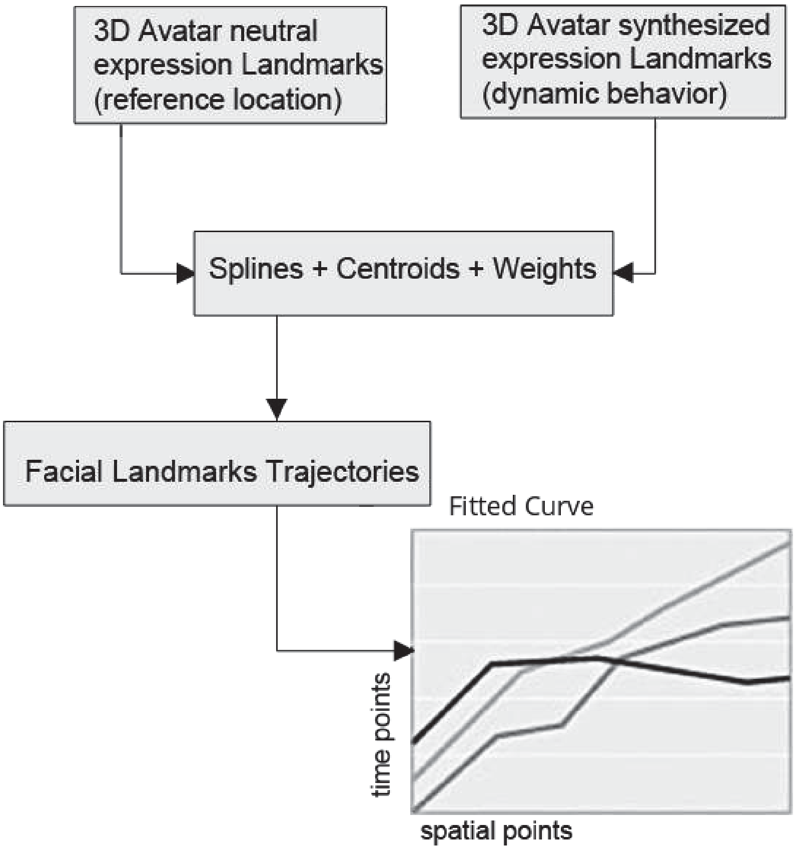

3D Avatar neutral expression Landmarks (reference location)
3D Avatar synthesized expression Landmarks (dynamic behavior)
Splines + Centroids + Weights
Facial Landmarks Trajectories
Fitted Curve
time points
spatial points

Figure 5.1: Model for 4D Region Data. Based on centroid coordinates from neutral to synthesized expression animation, a trajectory will be extracted using Euclidean Geometric Curves as B-Spline (Aldrich, 1998) that enable behavior analysis experiments. The Weight parameter refers to region influence obtained in Chapter 3.

point in a trajectory in synthesis process. In this way, for a trajectory referring to the synthesis of an expression E of degree n and Controller Points represented by the centroid of the region α observed as a time point $t(\alpha)$ as shown in (Eq.5.3):

$$t = p_0, p_1, \dots p_n$$

$$t = \forall p(\alpha) \in [0, 1] = (p_0, p_1, \dots p_n) p_i \geq p_{i-1}$$

$$t(C) = \frac{\sum_{v_{ir} \in R_r} \frac{|d_n v_{ir} - d_e v_{ir}|}{d_n v_{ir}}}{NV_r}$$

$$B(t) = \sum_{i=0}^{n} N_{i,k}(t) C(t) + W_{e(t)} \qquad \text{(Eq.5.3)}$$

Where for all control points p_i as centroid $t(C)$ and the knots in B-Spline consider the parameter W based on the value of influence extracted from PCA and FA analysis. The influence parameter W is taken into account in the animation process, where lower W correspond to less important regions that can be ignored in the expression synthesis, reducing the computational load.

The spatio-temporal region model was presented by Martin Erwig e Güting (1998) where a spatial points trajectory, based on temporal readings can be observed as a region if its spatial displacement is considered (Figure 5.2). When the shape of the curve changes, the region is expanded or retracted. The same concept is used in this work, once the centroids trajectories along the generated facial expression animation can be represented by dynamic curves, controlling the edges that connects the vertexes landmarks, producing the movement or deformation of facial regions as defined in the model presented in Chapter 3.

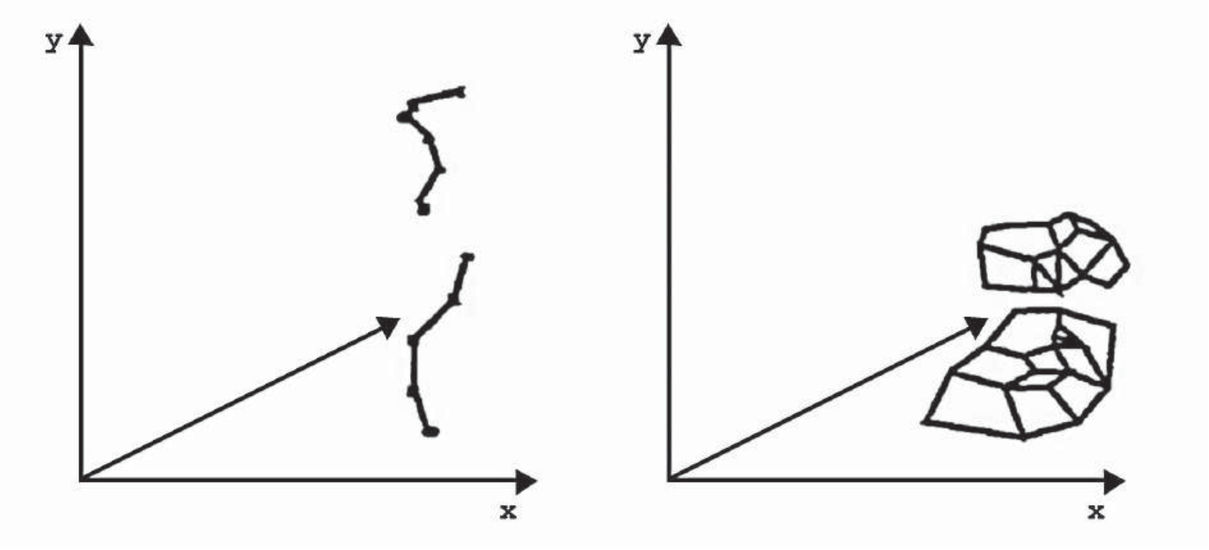


Figure 5.2: 4D Region (Martin Erwig e Güting, 1998). In the left, two curves referring temporal displacements of two spatial points. The coordinates of the curve represent spatial points (x-axis) and time points (y-axis). In the right, a region extracted from these curve. Based on the points defined in the curve, the related points are affected by the time displacement as a geometric influence.

Each 4D region was defined by the trajectory of its centroids (one for each region except for the mouth and eyes where the geometry is open, then having two centroids for these regions), observed temporally for each base expression, through its spatial coordinates, based on the concepts presented by Martin Erwig e Güting (1998), Dias et al. (2016) and Hongjian et al. (2016) as explained below. For the generation of the centroid trajectory, the Least Square Approximation algorithm will be applied in order to analyze more efficiently the influence of each trajectory in the synthesis process and the behavior of each facial landmark. Least Square Approximation algorithm adjusts data collected, for example, given $x_1, x_2, x_3, .., x_n$ in points of time $t_1, ..., t_n$, geometrically, it is possible to observe the linear relation between the temporal variables t and the observations x, fitting to polynomial to a given data Hongjian et al. (2016).

With the centroids curves referring to the region's landmarks represented by the deformation of the neighboring landmarks based on centroid displacement, it is possible to define 4D regions geometrically and calculate variations through polynomial interpolation for input coordinates. This model is compatible with that presented by Hongjian et al. (2016) and Mkrtchyan et al. (2016), allowing extension for any expression synthesis, including the interpolation of Ekman model emotions.

With this model, it will be possible to identify the behavior of the facial centroids through spatio-temporal facial regions. The regions as trajectory curves will be shown in the next section. Extracting the coordinates of the points temporarily will enable to understand the behavior of each centroid and thus the definition of strategies to optimize the animation of the synthesis of expressions.

5.2 4D TRAJECTORY EXTRACTION

Following the data model and the centroids coordinate extraction methods already defined in previous chapters, the next step of this research is focused on generating the spatial-temporal regions of the facial areas in the synthesis of the base expressions. The spatio-temporal regions are formed by the centroid spatial coordinates along a temporal trajectory of each area on the geometric surface during the synthesis process of the positive and negative expressions of the Ekman model, added to the weight parameter defined in the PCA and FA tests.

With the trajectories it is possible to apply a more complete analysis of the facial landmarks behaviors, supporting the decision about which methods to apply to an optimized synthesis process. The test parameters defined for the generation of the facial landmarks trajectories were divided into four stages: temporal interpolation of expressions, a division of the observation window of the frames, extraction of centroid coordinates, generation, and visualization of 4D Curves.

5.2.0.0 Trajectories Extraction and EDVA

The first stage consists of the extracted syntheses, including the expressions and facial region centroids for each expression in order to visualize the landmarks path in the interpolation of the expressions from the neutral state. Next, will be presented the observed window of the frames used and interest points for each generated curve, as well as the landmarks temporal visualization of there throughout the synthesis process. After the trajectories are generated, a process of optimization of the curves will be applied for final analysis.

The relevant coordinates occur in the transition between the neutral expression to the synthesis of one of the six basic emotions of the Ekman model since with these landmarks it is possible to observe the specific impact of these expressions on the 3D mesh. The centroid analysis in the interpolation of the expressions, also, can be extracted, generated and observed using the same procedures presented below, generating only a larger and more complex amount of data to be analyzed.

A regular animation runs at 30 frames per second (FPS), or 1800 frames in a minute, where each shot is a 2D render of the n-dimensional scene. In renders that use techniques such as scanline a manual parameters set are required for each frame, generating 1440 n values per minute for an n parameter model, with consistency between 24 30 FPS (Harold Whitaker, 2009). The authors define that in a 3D scene a frame is composed of points in high-dimensional state space, interpolated by splines. The splines are composed by coordinates for each point in space, having speed control and parameters for translation and rotation that can be plotted as trajectories.

Huenerfauth (2016) presents in his work a table that defines the number of frames used to register signals in signal languages, where, in order to register signals, regardless of its complexity, varying between movements that needs less than 10 frames to be recorded to the maximum of 78 frames. The ideal scenario for this work, once the facial expressions had their centroids displacements compared, is to keep a pattern of frames number by synthesis of expression, generating expressions in the same time interval of approximately 60 frames per second.

Goran J. Zajić e Reljin (2016) describe each frame as a featured vector which consists of color and texture information analyzed using a Multi-Fractal framework. Using a 4x4 regions partition, the authors perform an analysis in order to identify in frames of an animation characteristics that define whether the observed region is a computerized animation or an actual image based on the texture information. Each scene is sliced in blocks, where Feature Extraction algorithms and Analysis of Multi-Fractal Spectrum are used in order to classify each region.

For this work, a sectorized analysis process, similar to what Goran J. Zajić e Reljin used, was developed for the extraction of 4D regions. Each base expression animation has approximately 50 frames, separated by five main facial regions and three or four centroids associated with geometric mesh controllers. These facial landmarks are based on the Accurate Facial Landmarks proposed by Guo et al. in 2016 adapted to the MPEG-4 model chosen for this research, which defines vertices in the 3D mesh with the points of interest of the facial model defined in Chapter 3.

The trajectories of the Facial Expression Landmarks (FEL) defined in Chapter 3 were extracted using intervals of 60 frames for each Expression as shown in (Eq.5.4):

$$Traj = \sum_{i=1}^{n} Cent_{R_1}, [E_0, E_1] \qquad \text{(Eq.5.4)}$$

Where in a range between the neutral expression and the base expression $[E_0, E_1]$, the coordinates of the FEL are extracted by the region centroid $Cent_{R_1}$, defined in the Chapter 3, and their displacement. The trajectories can be organized graphically in 3 dimensions or in a simplified way in a Cartesian plane (Troy et al., 2016) as 2D curves where the time dimension information will be considered, representing the observed animation key-frame, and spatial, considering the centroid coordinates at the observed moment. Each centroid data can be compared to local deformation control points based on the synthesized expression as the data-driven concept where a point animation occurs based on coordinates extracted from an automatic input commonly used for facial animation (Li et al., 2016) (Caesar et al., 2016) (Gao et al., 2016).

The 4D region consists of the facial region centroids trajectory based on each of the base expressions. Thus, the test parameters for the trajectory extraction are built with a 2D curve representing the normalized centroid displacement in the expression synthesis process. The centroids displacements of the region define their behavior and influence for each expression. For a base expression, observing the 4D regions it is possible to point out which regions have the greatest geometric influence based on the three-dimensional Centroid Euclidean distance.

A controller was created for the vertices of the region $Cent_{R_n}$, in a range referring to the synthesis of the expression $[t_0, t_n]$ where the displacements are applied in centroid position and its coordinates stored in a vector. The algorithm was implemented with the expressions defined by keyframes and the centroids calculated from the vertices of the 3D model constructed in Chapter 3. For the graphical visualization of the trajectories were used methods from the OpenCV and Scypy libraries.

For each frame observed, with the coordinates of the Centroids, the three-dimensional Euclidean distance was calculated based on the spatial displacement. The process was repeated for each base expression, tabulating 50 C_r and T values for the Centroid of each region (the number of frames defined for the synthesis, excluding the 10 frames to return to the neutral position). For regions where symmetry does not occur as in the region of the cheeks, more than one centroid has been observed, as well as regions with openings in the mesh, such as the eyes and mouth.

Once the centroids coordinate points have been defined, Dummies controllers have been created in the virtual environment, which acts as rig structures associated with points in the geometric mesh. The coordinate values extracted in the reading are absolute, not considering whether the displacement of the vertex is negative or positive for a 3D axis, but its influence, or Euclidean distance, throughout the synthesis process. In this case, the Dummies differ from a traditional rig, for example using Bones controllers, since they do not influence distortions in the structure, but only make readings of the position of associated points, as the facial landmarks.

The trajectories were constructed using the NURBS curve technique, a vector-valued piecewise rational polynomial function. NURBS curve is commonly used to represent an action as a movement of a mechanical arm or the trace of moving objects (Aleotti et al., 2005) (Feng e Shen, 2017), and for this work was constructed using the coordinates of the Centroids as control points observed in a temporal stamp. Feng e Shen (2017) describes a trajectory as a sequence of tuples $T = [t_1, ..., t_n$, where each tuple t_i consists of a location $t_i.p$ and a time stamp $t_i.t$. In these time interval the object moves along the line segment between $t_i.p$ and $t_{i+1}.p$ from time $ti.t$ to time $t_{i+1}.t$. The author also presents a model to represent Time Synchronized Euclidean Distance of the trajectory control points. Figure 5.3 shows an example of the temporal trajectory of fifteen points, $[p_0, ..., p_{15}]$ and the representation of the positions of intermediate points by calculating the Euclidean Distance using the time parameter, that measure the errors through distances between pairs of temporally synchronized positions with a constant speed between p_1 and p_4.

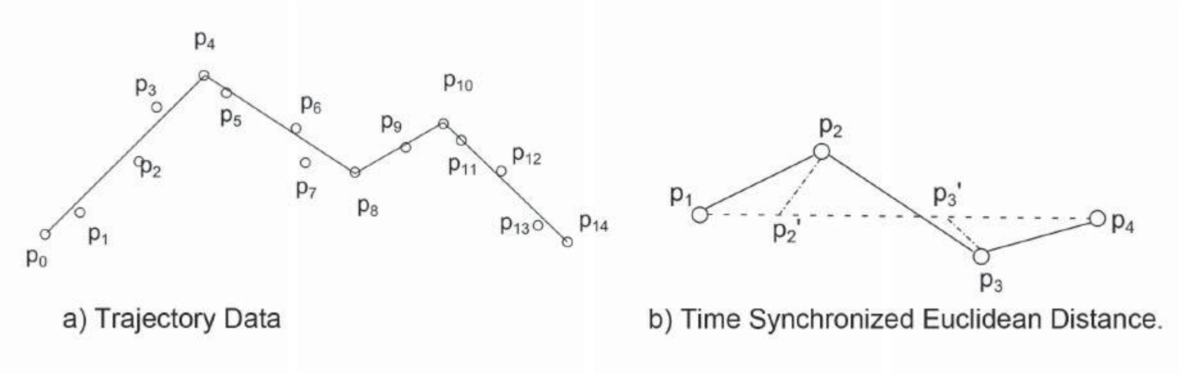


Figure 5.3: Data Trajectory representation in Temporal Stamp (Feng e Shen, 2017). a) Trajectory data for 15 points. b) Time Synchronized Euclidean Distance of the trajectory control points.

The curves shown in Figures 5.4, 5.5, 5.6, 5.7, 5.8 and 5.9 represent the displacement calculated by the Three-dimensional Euclidean Distance (3DED) of the Centroid coordinates $C_{E_1 1}$ observed at 50 ts (t relating to an element of the Keyframes vector for the emotion synthesis). Each line in images defines the Centroid trajectory by expression, the displacement projection can be considered the 4D data, once represent a geometric controller in a time slice. The graphs of representation for the multidimensional Euclidean distance used logarithmic scale in order to facilitate the visualization, for covering a large amount of values.

For the cheeks region, as well as in the mouth, more than one Centroid was extracted, since the calculations do not consider symmetry once the bases AKDEF and JAFFE have small variations in their tabulated values.

Zhai et al. used, in 2011, a Euclidean Distance Matrix Analysis (EDMA) to classify the shape of a face, based on the defined distances between the points of facial reference. Using Euclidean Distance Classification, Yashar Taghizadegan, Hassan Ghassemian presents a method for 3D Face Recognition using PCA applied in a 2D input using depth channel information (Yashar Taghizadegan, Hassan Ghassemian, 2012). Similarly, using Euclidean distance it is possible to identify landmarks geometric behavior through the temporal trajectories presented, pointing out the influence of each Centroid in the synthesis represented by the base expression.

Following, considering another option for trajectory analysis beside the plotted graphs, the values of the Euclidean Distance Matrix of centroids readings were calculated based on the displacement of the coordinate values extracted in the previous experiment. In order to understand the variance of centroids distances, Multidimensional Euclidean Distance was used, which for the Facial Landmark Centroids corresponds to 7 vectors with 50 dimensions. The data variation is calculated by the value of the last data subtracted by the first since the vector is ordered, and the variation of the complete Matrix calculated by a summation of the absolutes of the vector index by index for all values using N-Dimensional Euclidean Distance Variance.

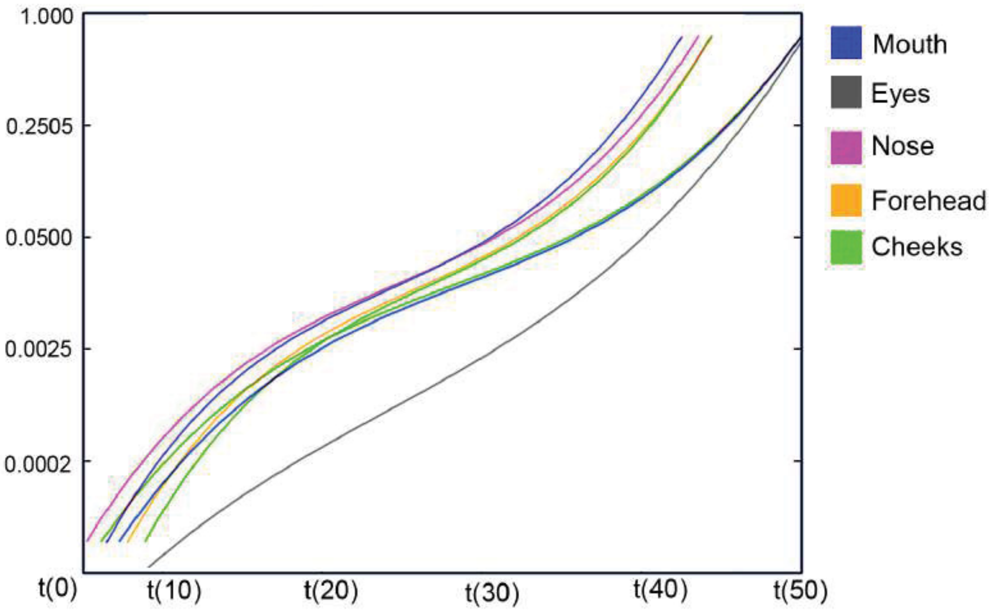


Figure 5.4: Centroid trajectories for Joy expression synthesis. It can be seen that the trajectories for the Centroids of the eye region, shown in gray, had the least significant displacement, with 0.000601 Euclidean Distance variance along the nose region, with 0.00025 of variance, as shown in Table 5.2. cheeks and forehead regions presented more expressive variance in their displacement coordinates with 0.01637 and 0.05269 variances respectively.

In the Figures 5.4, 5.5, 5.6, 5.7, 5.8 and 5.9 it is possible to observe the Centroid trajectories of the 5 facial regions for the base expressions. Graphically it is possible to observe which region has more spatial variation considering its geometric coordinates. The 3DED values of each Centroid reading are assigned to the matrix sequenced by the key-frame of the observed animation, as in Table 5.1 where EDV values of Centroid 1 of the mouth facial region are displayed and their overall variance value. Based on Euclidean Distances, Table 5.2 shows the Variance Analysis Matrix for each Centroid in the time slice of the synthesis of each expression.

Table 5.1: Euclidean Distance Matrix for $Region_1 C_1 \in [t_{Exp_{J}oy0}, t_{Exp_{J}oyn-1}]$ and the variance value. The values correspond to the three-dimensional Euclidean distance of the Centroid obtained from each frame of the animation for synthesis of the expression Joy.

EDV: 0.123670
ED Matrix:

0.02655852	0.03373305	0.04816728	0.06987239	0.09875764
0.1346078	0.17706123	0.22559191	0.27949768	0.33789951
0.39975368	0.46387903	0.52899871	0.59379208	0.65695312
0.71724534	0.77354995	0.82490075	0.8705044	0.90974512
0.94217889	0.96751821	0.98561153	0.99642122	1.0

Some values of the EDVA table are reinforced by PCA tests, such as the fact that the Nose region is more significant in negative expressions (such as Anger and Sadness) with a

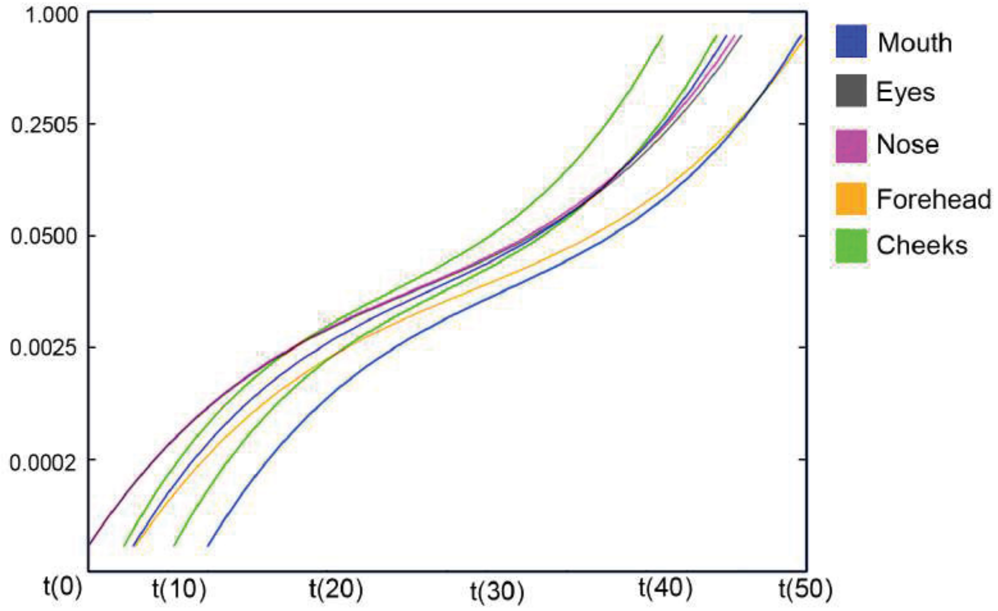


Figure 5.5: Centroid trajectories for Anger expression synthesis. For the Euclidean Distances Variance Analysis (EDVA), the regions with the greatest influence on the geometric mesh deformations did not present very great differences, highlighting the eyes, cheeks and mouth regions with 0.00236, 0.001104 and 0.00178 in the EDVA matrix (Table 5.2).

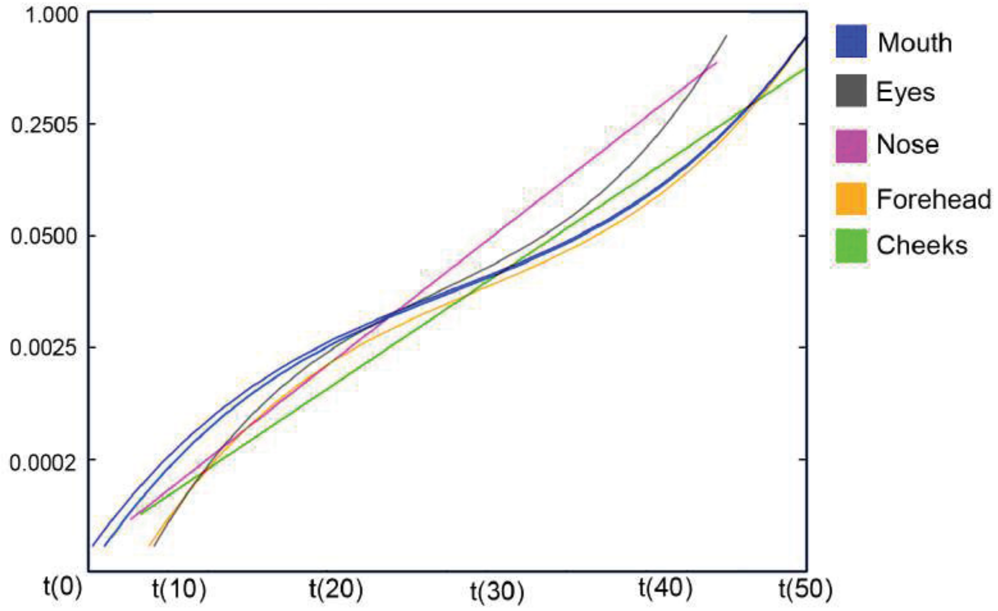


Figure 5.6: Centroid trajectories for Surprise expression synthesis. Based on the tabulated values in the EDVA matrix, the regions with the most prominence in the synthesis of this region were the superior ones, forehead, and eyes with 0.076342 and 0.001821. Whereas surprise may be a positive or negative emotion it makes sense that the attention in these two facial regions is greater than in the lower regions in the identification of this expression.

variance of 0.00236 and 0.004911 respectively, significant among the matrix average. For the synthesis of the expression Joy, the nose region has a smaller weight and can be ignored in the

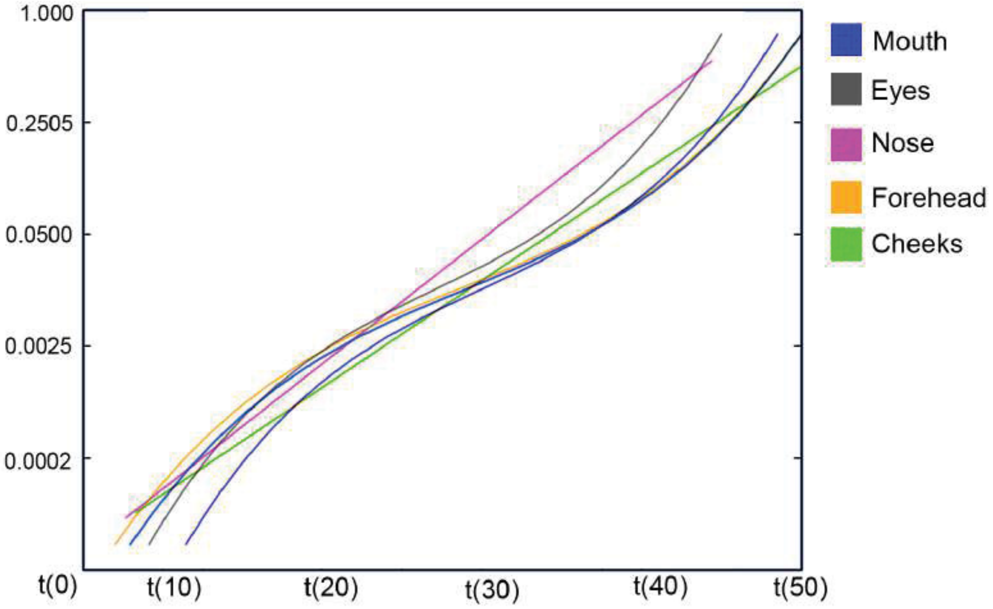


Figure 5.7: Centroid trajectories for Fear expression synthesis. The expression of Fear is considered more subjective than the other base expressions Lundqvist (1998), and it may be more difficult to classify locally values for this expression. Based on the blend-shapes applied in this work experiments, however, it is possible to point out the forehead, eyes and mouth regions as more expressive in the EDVA matrix, with 0.014019, 0.001821 and 0.010607 variance respectively (Table 5.2).

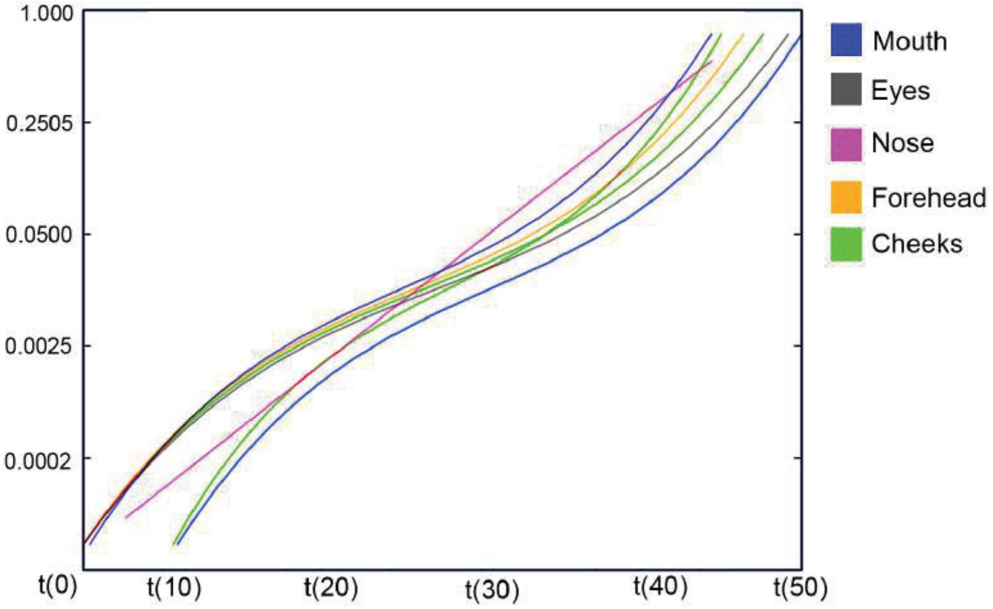


Figure 5.8: Centroid trajectories for Sadness expression synthesis. According to the values of the Centroid euclidean distance in the synthesis of this expression, the mouth region has more expressive displacements.

animation simplification process, and the mouth and cheek region can be considered the facial areas that define this expression with a variance of 0.123670 and 0.163710 respectively.

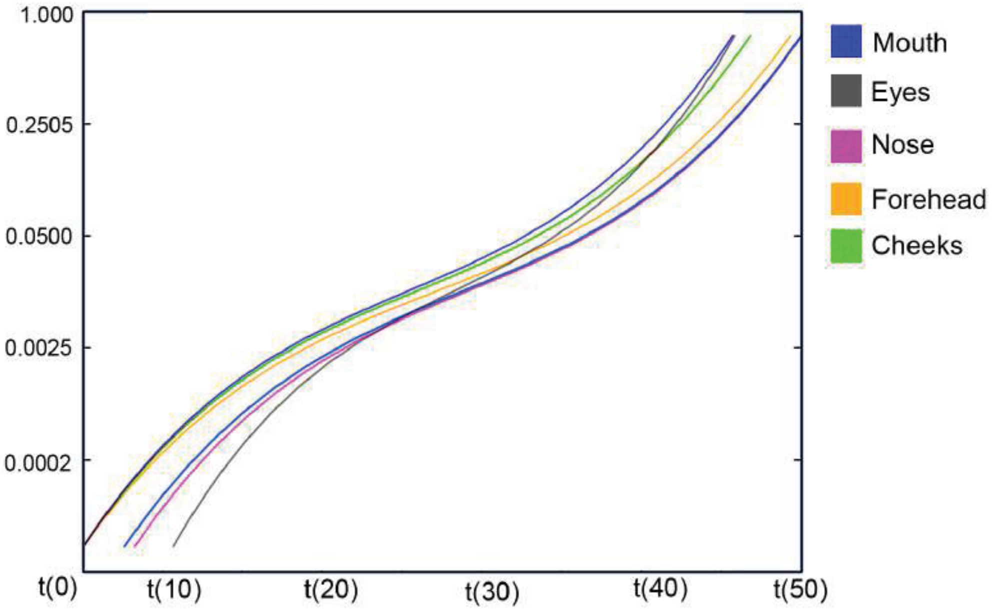


Figure 5.9: Centroid trajectories for Disgust expression synthesis. The values presented in the EDVA show that the mouth and forehead region have a greater geometric variance in the synthesis process with 0.418209 for the Centroid of the mouth region, one of the largest of the whole matrix, and 0.011317 for the forehead region.

Table 5.2: Centroids trajectory Euclidean Distance Variance Analysis values.

	EDVA values					
Centroid Region	Joy	Anger	Surprise	Fear	Sadness	Disgust
Mouth	0.1236	0.0011	0.0022	0.0106	0.1972	0.1820
Cheeks	0.1637	0.0017	0.0007	0.0007	0.0011	0.0039
Nose	0.0002	0.0007	0.0007	0.0007	0.0006	0.0007
Forehead	0.0526	0.0012	0.0763	0.0140	0.0041	0.0113
Eyes	0.0006	0.0023	0.0018	0.0018	0.0049	0.0013

The forehead region has greater variance in the positive expressions (Joy and Surprise) as well as the mouth region, reinforcing the variance obtained with the PCA Algorithm. For negative expressions, the mouth region has more significant variance values. The expression Anger, according to the analyzed matrix (Table 5.2) indicates that the regions of the eyes with 0.002369 and cheeks with 0.001781 are those that have greater variance in the displacements also reinforcing the values tabulated by the PCA test.

The Avatar mouth presented high values of geometric displacement variance in practically all expressions, highlighting disgust, with 0.1820 of variance and sadness, with 0.1972, which parallels the result obtained by the PCA algorithm, reinforcing their relationship between these expressions and facial regions. The cheek region has a greater variance in the Joy expression synthesis with 0.16371, and its lower values are in the expressions of Surprise and Fear, expressions that depend more on the Forehead region that presented high values of 0.076342 and 0.014019 respectively.

Factor Analysis results point out that the eyes and nose have fewer deformations in the geometric mesh, and can be considered, in general, the regions which exclusion or simplification less impacts the characterization of a synthesized expression. Conversely, the mouth region is the most important region, impacting more on the animation of the expressions followed by forehead and cheeks regions. These statements are reinforced by the analysis of variances in Euclidean distances matrix, where, again, the values applied to the nose and eyes are among the least expressive. In the same way, the highest values are in the mouth region, for all expressions being the most important region to reproduce a human emotion in a 3D avatar.

Table 5.3 presents the normalized variance data for the PCA and EDVA test relating the centroids of the facial regions and the base expressions. Although extracted values cannot be compared directly, due to the difference between the algorithms, the data presented for the influence of facial region on the synthesis of expressions reinforce the main regions that characterize the negative and positive emotions.

Table 5.3: Centroid displacement variance for Facial Regions in expression synthesis. The shown values were normalized and separated by the facial region and animated expression. For regions with more than one Centroid, the mean-variance values were calculated.

Principal Component Analysis normalized variance values.						
	Joy	Anger	Surprise	Fear	Sadness	Disgust
Mouth	0.626	0.005	0.011	0.053	1.0	0.923
Cheeks	0.829	0.009	0.003	0.003	0.005	0.020
Nose	0.001	0.003	0.002	0.003	0.003	0.002
Forehead	0.267	0.006	0.387	0.071	0.020	0.057
Eye	0.003	0.012	0.009	0.009	0.024	0.007
Euclidean Distance Variance Analysis normalized values.						
Mouth	0.791	0.702	0.970	1.000	0.400	1.000
Cheeks	0.800	0.953	0.000	0.673	0.307	0.667
Nose	0.376	0.080	0.000	0.095	0.091	0.000
Forehead	0.118	0.628	0.363	0.250	1.000	0.730
Eye	1.000	0.900	0.460	0.484	0.980	0.050

Through the histograms of the data shown in Figure 5.10, it is possible to perceive the proximity of the results of the geometric influence extracted from the PCA and EDVA tests. The spatio-temporal trajectories, however, allow the finer coordinate control of each dimensional axis by landmark gathered in the region, allowing complex interpolations and specific changes or simplifications in each synthesis. Those concepts are explored in the next section.

With the generated trajectories and the analysis of 3D avatar facial landmarks behavior, shown in Chapter 3, section 2, segmented by the regions based on their spatio-temporal coordinates in the expressions synthesis process, a method for simplification or optimization is shown in the next session. The coordinates of the centroids can be extracted in an external file by the parser system, shown in Chapter 4, and then modified to include new parameters or adapt the synthesis to other signal languages formal models.

Su et al. (2015) define the spatial relationship between trajectories and anchor points, equivalent to the facial control points, to fit the curves. The anchor points are used in this case as a mapping between the distance readings in order to interpolate or understand the behavior of the observed point over time. The authors use Dynamic Time Warp, an algorithm similar to PCA and FA explored in Chapter 3, over Euclidean Distance, besides other measure algorithms to plan new trajectories. Jain et al. (2009) use the same type of control point mapping, using landmark distances to project or to optimize the trajectory, based on the calculated behavior of the tabulated points.

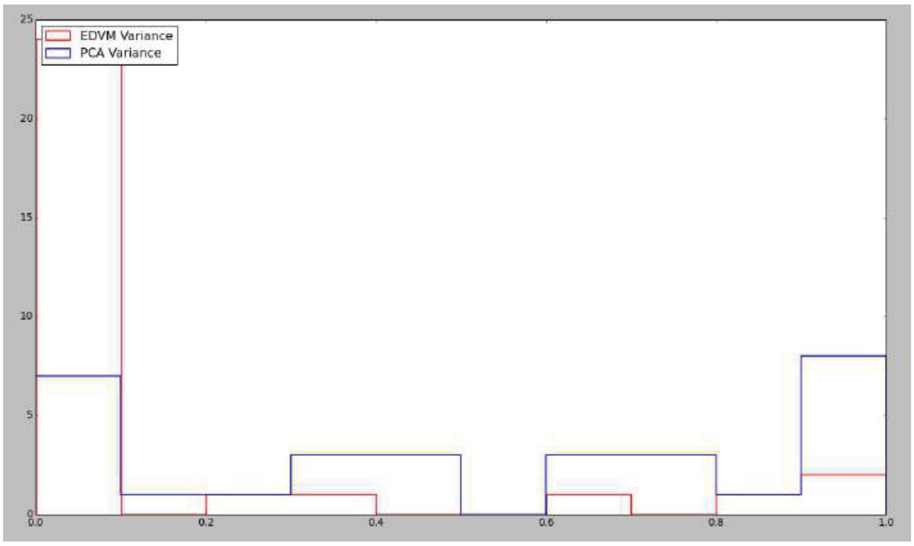


Figure 5.10: Histogram for Centroid variance values for PCA and EDVM based on facial regions.

The next section introduces the expression interpolation system using the concepts defined up to this point. The interpolation process allows to generate new expressions that use modulated parameters or intensities and can reproduce any necessary emotional variation controlled by facial regions through the model defined in this work.

5.2.0.0 Expressions Interpolation and Animation Optimization

Aiming to exemplify the simplification process, the 4D data of the 5 facial regions were exported for the synthesis of the base expressions, where for each region was defined vectors with coordinates referring to the temporarily observed landmarks. Figure 5.11 shows the process where, with AKDEF and JAFFE based blend-shapes, the coordinates are exported in the interval that refers to the time stamp for base emotions, in an external XML format file. Coordinates were simplified in order to optimize animations by excluding less relevant regions, or by simply interpolating the values of two or more expression to get a new emotion as shown by Szwoch (2015) in Chapter 2. The built-in interpolation process was used to generate a data set with outputs of new animations for the 3D avatar.

The main idea of this process is, with the coordinate export system and the matrix analysis of geometric displacements, spatio-temporal regions can be manipulated in order to simplify, and consequently, optimize, the facial animation process, defining which landmarks should be prioritized for each synthesis. In addition, this system allows you to parameterize the emotions synthesis in 3D avatars, allowing pointing new parameters and calculate other emotions or facial expressions animations, controlled by independent regions in a simplified way.

There are many approaches to simplifying or optimizing curves or data spatio-temporal data where intermediate points are calculated in order to obtain a simplified trajectory that respects the original format (Gloderer e Hertle, 2010) (T. Jusko, 2016), in the case of the 4D Centroids, the plotted geometric deformation from their coordinates. Li et al. (2014), from a weight parameter, simplifies curves by interpolating duplicated or less relevant edges shortening the number of controller points in the spline. The process was applied to create a data set with different outputs representing interpolated emotions or simplification of base expressions parameters, as well as to evaluate the operation of the data exporter system and new facial parameters application in the 3D Avatar.

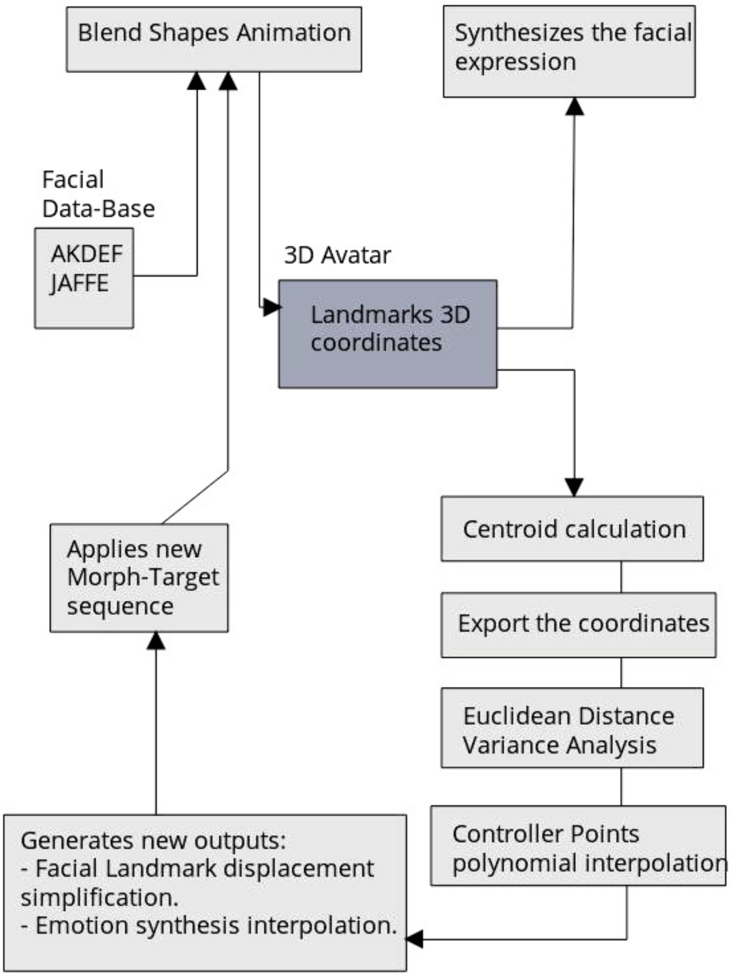

Blend Shapes Animation
Synthesizes the facial expression
Facial Data-Base
AKDEF JAFFE
3D Avatar
Landmarks 3D coordinates
Centroid calculation
Export the coordinates
Euclidean Distance Variance Analysis
Controller Points polynomial interpolation
Applies new Morph-Target sequence
Generates new outputs:
- Facial Landmark displacement simplification.
- Emotion synthesis interpolation.

Figure 5.11: Generation of new outputs for the synthesis of expressions in the 3D avatar. New simplified animations, or expressions calculated by the interpolation of base emotions, can be constructed.

Using the parser system applied to the 3D avatar, average control points were defined using Newton's polynomial interpolation, calculating coordinates for midpoints, in key-frames time-line, by limiting the calculation of the landmarks displacement during the animation. Interpolation allows constructing a new coordinates data from a discrete dataset of previously known information, in the case of the 3D avatar, the spatially tracked coordinates. Through the polynomial interpolation, a function was constructed that finds intermediate values for the coordinates of the Centroid, calculating a new position for the facial Landmarks.

The Newton Polynomial Interpolation Algorithm was used in the process. Through a list containing the values of the coordinates separated by sequential readings of the expressions, the simplification method of the 4D regions was implemented. The readings must contain values in the three geometric axes for each key-frame, defined by 50 frames for each expression, plus 10 frames to return to the neutral state, exporting a different list for each landmark. The landmarks are further grouped by facial region, in order to parameterize the simplification process by datasets. The parser generates an XML file with float value lists, and Newton's polynomial interpolation numerical method is applied, looking for intermediate values in the input. The tabulated values have the dimension of $y[T_a, T_b]$, $x = [Cent_0(x_0, y_0, z_0, x_1, y_1, z_1, ..., x_n, y_n, z_n), .., Cent_n]$, referring to the temporal slice defined for the expression in $[a, b]$ interval.

With a list of intermediate values u, vertexes spatial displacements are simplified generating a new spatio-temporal trajectory. With the exported coordinates it is possible to further soften the animations by decreasing the variance in the displacements or reducing the computational cost by excluding 4D regions with a lower w parameter. The value of w represents the weight extracted from the EDVA, PCA and FA results, which define the weight relationship between the expression and the facial region. Once the value of an intermediate point u is calculated, its coordinates can be inserted in Centroid values list of the blend shape by the distance between Centroids and the data matrix using the Function that calculates the distance between C_u and C_j in $[C_0, C_{n-1}]$.

The Terror expression is synthesized by intensifying the Fear expression values, classified by Lyons et al. (1998) and Lundqvist (1998) facial bases, just as the expression of Ecstasy that is obtained by intensifying the values of Joy expression. The opposite path can be obtained, for example, to represent Annoyance in a 3D avatar, just need to soften the expression of Anger following the (Szwoch, 2015) emotions model. These conditions are possible by interpolating all facial regions in the synthesis by rendering a more contained expression with smoother geometric displacements. In the same way, it is possible to interpolate base expressions as in the case of the Apprehension expression, where the values of the expressions Fear and Surprise are processed. Joining parameters of Joy and Fear expressions in different intensities we have expressions like Love or Neutral Submission, varying the values of each base expression and its regions. All facial variations can be reproduced by the 3D avatar manipulating the landmarks coordinates values of the 5 regions, based on the behavior of the base expressions.

The Figure 5.12 exemplifies the interpolation process for the avatar where, in the upper part of the image, the neutral expression is represented in the AKDEF facial base, used to create the blend shapes, and the avatar with the landmarks in the initial position. Following there are the outputs for the expression of joy, in the left, and fear, in the right, with all the parameters with the standard intensity calculated based on the databases AKADEF (Lundqvist, 1998) and JAFFE (Lyons et al., 1998). The coordinates were exported using the parser system where Newton's Polynomial Interpolations was applied for new control points in the facial regions, varying the intensity between the two base expressions. After this process variations of the standard synthesized expression are obtained.

Following, at the bottom of Figure 5.12, two 3D avatar syntheses are shown where the values of the facial regions have been interpolated to represent new expressions. The first variation presents a neutral synthesis maintaining 30% of the region 4 and 5 intensity of the Fear expression and 15% of the values intensity of Joy expression for the same regions, where a submissive neutral expression can be obtained. Then for the expression Love, the inverse was applied to maintain the intensity of the joy expression greater than the expression fear.

In the same way, local variations of expressions can be obtained by maintaining negative and positive aspects, intensifying or eliminating sectored values in order to get specific expressions or emotions. Associating exported landmarks with pre-defined controllers in the 3D engine, the process can be used for any variation by changing the context of the emotion as needed.

It is possible to apply the facial model for signal language parameterized in a simplified way in other 3D avatars or systems adapting the facial regions to the Centroids and local landmarks. The process of parsing and interpolation of expressions works with hierarchical parameters as, in general, sign language synthesis systems in virtual environments work. In addition, while evaluations and applications use Brazilian Sign Language as a motivation, emotions and facial expressions works for a conversation in any sign language and can be used for systems with 3D Avatars in general.

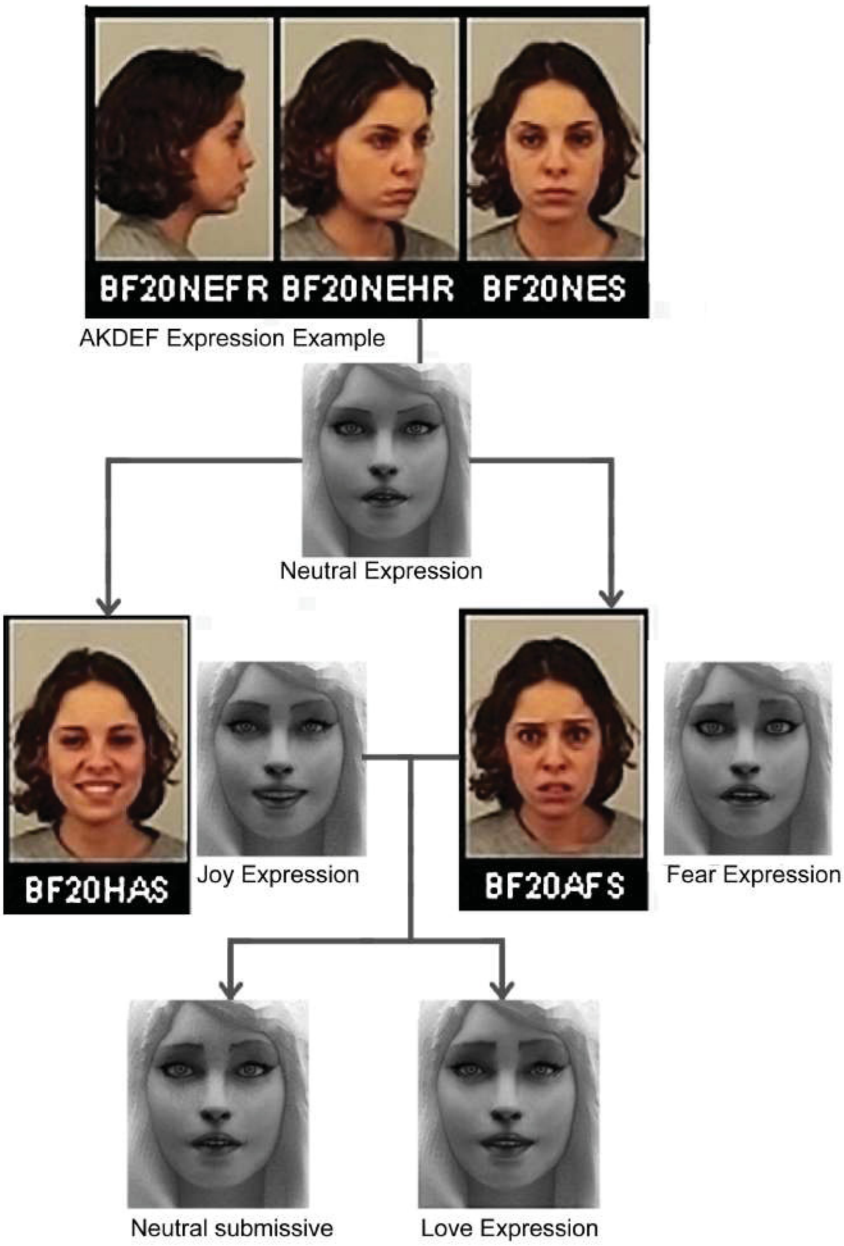


Figure 5.12: Example of expression interpolation using original blend shapes from AKDEF (Lundqvist, 1998) and JAFFE (Lyons et al., 1998) facial datasets. As seen in Chapter 3, the nomenclature used in the AKDEF dataset follows the interpreter identification (BF20 in this example), the expressions (NE for neutral, HA for joy and AF for fear) and position (S for frontal, HR for perspective and FR for side). From the expressions Joy and Fear, alternating the intensity of the spatio-temporal values of the facial regions in the two base expressions, variations can be achieved as the expression of neutral submission or love, as implied in Ekman's model (Szwoch, 2015).

With the presented system, a dataset can be generated with variations of facial expressions applied to the 3D avatar, where it is possible to evaluate alongside users of Brazilian Sign Language users the quality of the syntheses extracted according to the behavior observed for each facial region in the synthesis of the expressions and emotions.

The next section evaluates the generated outputs in the experiments applied to BSL language users, in order to observe the quality of the interpolated expressions and to analyze, through these data, the results obtained from the algorithms and methods developed throughout this work.

5.3 DATASET ANALYSIS AND FACIAL EXPRESSION VALIDATION

This section presents the application of a questionnaire in order to integrate the deaf people community with the results and the facial model obtained. Based on the 4D trajectories obtained in the previous experiments, some considerations can be made in order to optimize the synthesis process. Since the trajectory of each centroid can be exported using the parser system, it is possible through linear interpolation to define new control points based on the results of the PCA and FA algorithms application. These new inputs improve the synthesis process by adapting new parameters or by facial expressions interpolation by changing the behavior of the centroids based on the 4D region's influence.

In order to develop the new outputs to be evaluated, some protocols were defined following the parameters of the facial model MPEG-4 + Facial Regions. Next, the complete experimental protocol for the facial expressions synthesis output dataset are shown:

- For each new output, the coordinates of the Facial Centroids will be exported in an external XML file, calculating the interpolation of the average control points in order to reduce the number of displacements returning simplified spatio-temporal trajectories for the facial Landmarks.

- New outputs for each base expression for the Neutral expression will be interpolated, as follow: a) Complete synthesis following the MPEG-4 + FR model, and the blend-shapes based on the AKDEF (Lundqvist, 1998) and JAFFE (Lyons et al., 1998) facial databases. b) Simplified synthesis, eliminating the parameters with low influence according to the *w* parameter calculated from the geometric displacement tests developed in Chapters 3 and 4. c) Complete synthesis by interpolating/simplifying the weak parameters, in order to reduce their displacements but not eliminating their participation in the animation of the new expression. d) Synthesis with interpolated base expressions in order to generate a new expression, presenting some of the variations of the emotions model (Szwoch, 2015).

- Each new output will be built using the Blender Game Engine (BGE) tools, the Python programming language with the Opencv, Numpy, and Scypy libraries, all of which are open source software's and run on an NVIDIA GeForce GT 630M graphics card with 8Gb RAM using OpenGL 4.4. The same resources were used in the previous experiments of this work. It is important to note that although the experiments are executed in a virtual environment, the 3D avatar was created with a simplified low-poly mesh, and the computational cost of the whole operation is small, taking less than 1 second to export the coordinates, plotting of trajectories or to go through the vertex coordinates of the model in real time.

Based on the presented protocol, a dataset was constructed using the spatio-temporal data export process, classification and interpolation of trajectories, where we have examples, from the expression Neutral, of simplified or interpolated expressions. These variations are defined by region, intensifying the expression through the coordinates of each specific region in order to obtain a new expression. The Figure 5.13 shown interpolated and simplified expressions output. The constructed dataset have 22 synthesized outputs concerning 06 objects for each base emotion (Joy, Anger, Fear, Surprise, Sadness, Disgust), 08 objects for simplified base emotions (the previous 6 excluding the less influence regions by the parameter *w* indicated by the result of NEDeF, plus Optimistic Neutral and Pessimist Neutral) and 08 objects concerning interpolated secondary Emotions pointed out by Plutchik's Wheel of Emotions (Love, Remorse, Disapproval, Aggressiveness, Apprehension, Annoyance, Terror, Serenity).

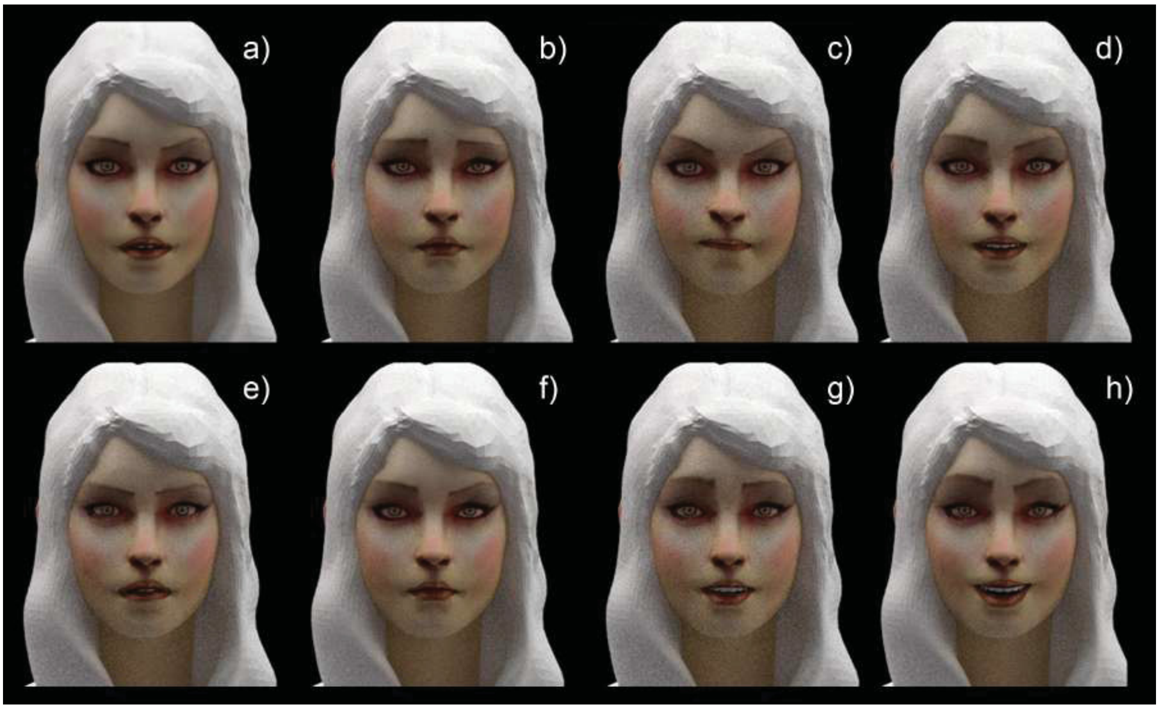


Figure 5.13: Dataset examples of interpolation and simplification of base expressions. From the neutral expression (a), simplified base expressions are presented. In the first case (b), the expression of sadness is generated by excluding the centroids of the nose and cheek regions, followed by the expression of anger (c), which was simplified by calculating polynomial interpolation using the half of the values tabulated from the blend shapes. The next images present interpolations of the base expressions, as neutral / optimistic (d) that is generated applying the displacements of the expression joy (h) in little intensity, being able to generate expressions like love (g) applying low intensity in more of an expression, in this case, joy and anger. Applying low intensity in anger and disgusting is obtained the expression of disapproval (f) or annoyance (e).

Using the created dataset outputs, a form was made available online to evaluate the system and its synthesis process, where some information of the synthesized expressions can be observed and compared to the behavior extracted from the centroids spatio-temporal trajectories. The form does not aim to evaluate the user in any sense, or to involve user person in any way in the validation process, but to support the mathematical evaluation of the generated expressions in the 3D avatar rendered with different parameters and interpolations of values in their regions. The application of the form enables a first contact of the techniques developed for this research experiments and possible users of applications that incorporate the systems of facial animation proposed. The complete form can be found in Attachment 1.

The information applied in the form was separated into two sections. The first one receives, anonymously, how many feedback's were from fluent speakers or users of Brazilian Sign Language, useful information for the statistics of the second section, which presented some outputs with variations on different expressions through the 3D avatar. The application of the form received 30 feedbacks, with 40% of BSL fluent speakers corresponding to 12 responses, 50% for users knowledgeable of the BSL language but non-fluent speaker and 10% non-BSL users. Was reported that 46.7% of the respondents already used software with 3D avatars as sign language interpreters, 10% already used software that uses 3D avatars but not in the context of sign languages and 43.3% never used a software with these characteristics.

Following, 86.7% of respondents consider facial expressions a very important feature in a sign language conversation, 10% consider it important and 3% consider it less important. These initial information add a greater weight to the observations of the form second section where the synthesized expressions will be validated since the respondents, for the most part, belong to users of the BSL language with some familiarity in 3D avatars sign interpreters.

A collection of 21 synthesized expressions outputs were presented by grouping direct combinations calculated by the parameters of the AKDEF and JAFFE facial bases, interpolations of different base emotions, altering their intensities beside simplifications of expressions excluding regions of lower weight, as presented in the previous sections. The second section of the form aimed the observation of the synthesized expressions where the respondents pointed out which emotions best represented each synthesis output, in order to validate the computational method developed and the final rendered emotion.

Was included in the form 14 outputs representing the base expressions of the Ekman model with 06 images being generated considering the displacement of all facial regions and 08 using the simplification process. The outputs representing the base expressions had only two errors of the 30 responses, being positively recognized by the users. The simplified base expressions presented 80% of correct answers, the basic expression Disgust received 4 responses as "Unidentified" and the expression Fear, with had 2 incorrect answers. We also included 07 outputs representing secondary emotions using the interpolation method. In these images 40% of the answers did not correctly identify the emotions summing 12 of the 30 responses. The result in these cases is still positive but contained a greater number of errors because the secondary emotions did not exist in the chosen bases JAFFE and AKDEF, and were generated only through polynomial interpolation.

The Table 5.4 presents Spearman's rank correlation coefficient and significance applied to the data extracted from the feedback obtained in the application of the form. The data below were organized in order to relate the importance of the parameters used to define the outputs in the user's responses. Between the 21 outputs included in the form (presented in Appendix 1), 8 of them presents simplified base expressions ignoring less influential regions based on previous tests, and 7 outputs present interpolation of two or more base expressions.

Table 5.4: Spearman's rank correlation coefficient applied to analyse parameters in the applied formulas. The presented values vary between 1 for direct relation between the parameters and -1 for inverse relation. 0 represents no relation. The parameters presented are: 'Base Exp.', for base expressions, 'Interpolation' and 'Simplification' for outputs of these categories, 'NDA' for unidentified expressions, 'Error' and 'Hit' for the number of errors and hits by output image.

	Base Exp.	Interpolation	Simplification	N.I.	Error	Hit
Base Exp.	1.00000	-0.75310	0.19113	-0.33467	-0.09853	0.19263
Interpolation	-0.75310	1.00000	-0.30900	-0.09937	0.13459	-0.06527
Simplification	0.19113	-0.30900	1.00000	-0.38683	0.11440	0.04804
N.I.	-0.33467	-0.09937	-0.38683	1.00000	0.15996	-0.51107
Error	-0.09853	0.13459	0.11440	0.15996	1.00000	-0.90577
Hit	0.19263	-0.06527	0.04804	-0.51107	-0.90577	1.00000

As expected the simplifications are directly related to base expressions, with 0.19113, with the base expressions being the most related to the correct answers in the identification of the expressions. This is reflected as being main emotions and due to the parameters well defined by the facial bases AKDEF and JAFFE, the simplified expressions have a positive correlation with the correct answers, appearing with 0.04804 in the Table 5.4. The relationships between the "Base Exp." and "Simplification" variables are not 1.0 because the correlation coefficient considers all the base expressions and simplifications not being able to relate one expression generated with the complete synthesis with a different simplified base expression. The positive relation, however, confirms the relation between these variables.

The image 16 presented in the form, for example, presents a simplified Surprise emotion synthesis limiting the intensity of two regions, had 23 hits between the 30 responses, as well as the image 17 that presents the expression of anger also simplified had 16 hits followed by the next image of the form with 15 hits for simplified expression of surprise by the exclusion of less influential parameters. These examples further divide responses from similar interpolation options (such as anger irritation) and it is therefore correct to say that the identification of these expressions was positive with a high hit rate. The Table 5.5 presents the cross-tabulation of the hits in relation to the simplifications. It is expected that, if these secondary emotions were mapped onto a base of 2D images with facial emotions, tracking the landmarks would result in more distinct and therefore more recognizable interpolations. At this point, the mean displacements of the base emotions suggested by Plutchik's wheel of emotions were considered.

Table 5.5: Cross-Tabulation between the Hit and Simplification parameters in the responses obtained from the application of the form. From the 30 identifications per image, it is possible to notice that the values for simplified expressions were high and constant. The simplified base expressions presented 80% of correct answers, validating the simplification and synthesis process.

Hit	4	5	7	9	11	13	14	15	16	17	19	20	21	22	23	26
Simplification																
0	1	1	0	1	1	1	0	2	1	1	1	0	2	0	1	1
2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

The images with representations of expressions interpolated as distrust or love had 0.13459 of correlation with the incorrect identifications, which is justified because they are combinations of intensity of base expressions that divided the answers between the correct one for each case and options of the base expressions that generate that interpolation. Another parameter that reinforces this deduction is the high relation of unidentified expressions with the errors, appearing as 0.15996. Responses in interpolated expressions tend to be more fragmented due to the subtle difference between interpolations and base expressions, which can be minimized when associated with the context of the message or the gestures associated with that expression when applied in message transmission.

Most of the incorrect responses do not represent erroneously recognized emotions but unidentified synthesized emotions. This occurred especially in the secondary emotions because they closely resemble base expressions. Intensifying the syntheses could make them more distinct, being a strategy very used in avatars with cartoon visual, and even by human interpreters. An alternative would be use in the blend shapes mapping, facial database containing the emotions represented with greater emphasis. The cross tabulation of the responses obtained for the parameters of "NI" (for unidentified expressions) and the "Error" is presented in Table 5.6.

Table 5.6: Cross-Tabulation between the Error and NDA parameters in the responses obtained from the application of the form.

Error	3	4	5	6	8	9	10	12	13	14	15	16	17	20	23
NI															
0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0
1	1	0	0	0	0	1	1	1	0	1	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0
3	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
5	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

In Table 5.7, the answers for interpolated expressions and number of correct answers are presented, where it can be observed that, even with a negative correlation in the Spearman coefficient calculated values, we find about 5 interpolation outputs with more than 16 correct answers of 30 answers each, a very positive number in the recognition of the most complex expressions. This means that even in cases of greater difficulty in recognizing the generated emotions the number of hits is still greater than that of errors. At this point it is safe to affirm that the synthesized expressions and the process of generating emotions in the avatars is functional and could add information and context to the purely manual messages of current systems of sign language generation using 3D avatars. The process can be improved in future works integrating morphological parameters of signal languages, but it is a regional constraint since each sign language has its own morphological and syntactic rules.

Table 5.7: Cross-Tabulation between the Hit and Interpolation parameters in the responses obtained from the application of the form.

Hit	4	5	7	9	11	13	14	15	16	17	19	20	21	22	23	26
Interpolation																
1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
8	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

The positive correlation between the basic expressions and the simplifications with the correct ones justify, then, the use of the proposed and implemented system, being a viable model in the first moment for the parametrized representation of facial expressions in signal language systems, optimizing the process by the simplification of values of regions with less relevant behavior in the synthesis process. For interpolations, it is necessary to define values on the facial bases for these emotions, as well as to include the manual parameters and message context. In this way, with more information available, the representation of the signals through 3D avatars has a quality gain helping the user to define the tone of the information and adding complex emotions to the signals.

The results are good fitting with the research proposal and he goals aimed once all the emotions synthesized through the process had a high degree of recognition by the users. The incorrect feedbacks will still help in the proposal of strategies for future improvements in the system. At the computational level, the implemented model allows local changes in the parameter values for the synthesis of each expression, isolating based on the behavior of the centroids temporarily, less influential regions which offers optimization in the process, also allowing the increment of new parameters and the generation of expressions of different intensities and complexities, manipulated by independent parameters. The results of the application of the form complement the geometric validation with the trajectories presented in the previous sections, contributing to define the behavior and influence of the facial regions in the synthesis of the expressions applied to the centroid model adapted from the MPEG-4 FP format.

5.4 CONCLUDING REMARKS

This chapter presented the parser system for spatial-temporal data extraction of 3D avatar facial Centroids. Also, experiments were developed defining the behavior of facial regions Landmarks, which allows optimizing the synthesis process and enables the application of this model for 3D avatars facial expressions interpolation for use in signal language synthesis systems. A dataset with outputs for base emotions was constructed, presenting objects generated with the proposed simplification method through the experiments of geometric influence by displacement of landmarks (NEDeF), and a model for polynomial interpolation of emotions in order to generate secondary expressions. A validation process of the constructed system was also presented, with BSL deaf community users, which validated the model by analyzing the outputs constructed for the dataset. All experiments used a 3D avatar mapped with the facial model presented in the Chapter 3, which validates the use of the whole process in a practical environment as a BSL landmarks facial model.

6 GENERAL DISCUSSION

This chapter presents a discussion about the main contributions and results obtained in this research.

6.1 3D AVATAR FACIAL DATA MODEL AND EXPORT SYSTEM.

Initially, in order to obtain the facial parameters definition, a model was developed in the Chapter 3 for use in 3D avatars interpretive of signal languages. The geometric facial points of interest, or landmarks, used MPEG4-FP model as the basis. A model that separates facial points into 5 major regions was proposed (Forehead, Mouth, Nose, Eyes and Cheeks), and blend shapes were then applied for each expression using the AKDEF and JAFFE facial image bases. Each landmark of facial control was defined using the Neutral Expression Deviation Factor (NEDeF) model proposed for this work, which allowed, using the 6 base expressions (Joy, Anger, Surprise, Sadness, Fear and Disgust), to apply the Principal Components, Factor Analysis and Explained Variance Ratio (EVR) in order to verify the relation of the facial regions and the basic expressions.

Interesting results were extracted from these experiments that define the importance of each region for each base expression, being, in normalized values, 0.79 for Forehead, 0.0 for for Eyes, 0.85 for Cheeks, 1.0 for Mouth and 0.13 for Nose region, which can be compared, for example, with the influence of a region for all the base expressions as in the case of the Mouth region, the most influential that appears using the NEDeF model with variations normally between 0.7 and 1.0 (high values in comparison with the other regions).

These first experiments were fundamental to the solution of the main weakness found in the bibliographic review presented in Chapter 2, which is the lack of a formal computational model for facial points for use in Signal Languages. With the NEDeF model, facial points of interest can be analyzed and parameterized, as well as described through computer models such as CORE-SL, facilitating the integration with parameters already mapped in sign languages such as BSL.

Aiming this integration it was proposed the use of a hierarchical data exportation method for the facial points compatible with the descriptive model used to define BSL manual parameters, based on the XML format, exporting the Centroid coordinates calculated by the NEDeF model, as well the parameters regarding the expression, with regions and facial information.

6.2 SPATIO-TEMPORAL CENTROID BEHAVIOR.

With the defined facial model, methods were developed to analyze the facial landmarks during the synthesis process, with the objective of evaluating the outputs of expressions based on the NEDeF model and proposing a method of expressions interpolation for the generation of other facial expressions besides the 6 base expressions. For this purpose, spatio-temporal data concepts were applied in conjunction with the NEDeF model using Euclidean Distance Variance Analysis (EDVA) in order to compare the influence of each region with the results obtained in Chapter 3 using the data correlation algorithms. The displacement of a Centroid can be observed by constructing a trajectory by reading its coordinates along the animation referring to facial expression. These spatio-temporal trajectories help in visualizing the influence of a facial region for a base expression, allowing to understand which regions are more relevant in the interpolation of more complex expressions.

Following, simplification techniques were applied in less affected facial regions, using polynomial interpolation in control points associated with 3D mesh and animation rigged structure. In this way, a data set was constructed in order to generate different outputs with interpolated expressions based on the variations presented in the Plutchwick emotions wheel. An online form was made available with the intention of reinforcing the obtained results from the implemented methods, where the variations of interpolated emotions and the fidelity of the simplified expressions were verified in a first moment.

40.6% of the validation responses were fluent in BSL, with 48% being knowledgeable non-fluent in the language. Approximately half of the users had already used some Sign Language software with a virtual interpreter, and 84.4% of users responded by considering very important facial expressions in a sign language conversation. The Spearman's rank correlation coefficient algorithm was applied to the answers in order to identify the relation of the errors and hits with the presented outputs. The basic expressions, even when simplified, had a high hit rate, motivating the inclusion of expressions in avatars interpreters of sign languages. The interpolations, since they were not mapped on the AKDEF and JAFFE bases and were, therefore, built based only on Plutchik's Wheel of Emotions, had a higher error rate, which suggests that targeted studies can be done to map the interpolations of emotions using the region facial model and the NEDeF model calculations.

The NEDeF model behaved satisfactorily for a definition of facial parameters for sign languages, being able to be incorporated into models such as CORE-SL and improved in future works, defining more precise displacements for interpolation of expressions, and joining the expression to a sequence of signals, applying context to the message interpreted by the 3D avatar. The next Chapter presents the Conclusion and suggestions for future work.

7 CONCLUSION

In order to represent a sign language computationally, parameters that follow a formal descriptive model and that can be applied to a virtual interpreter since sign languages are required. Although there are effective models in the literature that define the parameters necessary to represent manual signals, facial expressions are still an unexplored field for Brazilian sign language, BSL, although extremely important for the context of the transmitted message. This is the major problem addressed by this research, which proposed to identify and apply a model to define facial parameters computationally and compatible with the formats used to represent hand and arm signals.

As the main contributions of this research can be pointed out the definition of a computational data model specific to facial points, using MPEG4-FP model alongside a proposal of facial regions. This model was used in order to define the behavior of the facial expressions using a novel geometric landmark model called Neutral Expression Deviation Factor (NEDeF) extended for generation of spatio-temporal trajectories.

The behavior of the expressions with geometrically independent regions, allowed the mesh analysis of 3D avatar interpreter and the identification, through the application of data correlation analysis algorithms, such as Principal Components Analysis, Factor Analysis and Euclidean Distance Variance Analysis, defining not only the relation of the expressions with the facial points as enabling interpolation strategies and simplification of facial expressions for 3D avatars. This is important since less complex synthesis processes are more interesting to justify the integration of facial expressions into sign language representation software.

A dataset with simplified base expressions and expression interpolations was created and validated with BSL users where it was possible to point out the validity of the models and methods developed for this work, besides pointing out parameters to be worked on in future researches.

During the experiments, some problems or limitations could be identified. The facial parameter data export system proposed in this research is based on the CORE-SL model, that does not have integrated facial parameters, relying on future work for its implementation and integration in the format used to represent manual signals. Also, the experiments were done using a 3D avatar constructed for this research, being necessary the application of tests using other geometric meshes in order to compare the influences and behaviors of the facial regions and expressions.

The major limitation, however, is regarding interpolated expressions. In the validation stage it was possible to notice that for a more precise synthesis the ideal scenario is that there is a base that contemplates interpolation so that the blend shapes will be more distinct and generate more identifiable outputs. Another area to be considered in the next works is the inclusion of BSL morphological elements (Felipe, 2007), increasing the already mapped emotions with linguistic models defined specifically for the generation of signals through facial shapes, which include new areas such as the tongue (Felipe, 1997), (Iatskiu et al., 2017), (Felipe, 2006).

The focus of this research was on the synthesis of emotions by adding facial parameters to 3D avatar but can be extended to facial signals, since in BSL, as well as many other sign languages, it is possible to represent signals by specific facial expressions. Still, future work can map the iris movement in the eye, which can define information in specific signs, or the of the message.

It is understood that, through the parametrized facial model and the synthesis system, more specific processes can be implemented and optimized and fluid animations can be obtained in future works adapting the parameters and the geometric models. Furthermore, it is expected that representations of more complex expressions can be added to systems that use virtual sign language interpreters, using concepts and techniques presented in this work.

With the facial model defined in this work and the methods for facial landmarks behavioral classification for expression synthesis, it is possible to automate facial animations in new models of 3D avatars sign language interpreters. Also, the spatio-temporal data model can be applied to gestural parameters, aiding in the behavioral analysis of other Signal Languages parameters.

In this work, it was demonstrated that through a parametrized facial computational model it is possible to improve the transmission of messages in sign languages using 3D avatars with few impacts in the computational cost to generate animations of base expressions. It is expected that in future works, these methods will help the 3D avatars interpreters of BSL to present a more complete representation of the signals messages transmission.

The results obtained until the end of this research are an initial impulse that could be the starting point in the integration of facial expressions for sign language formal representation through 3D avatars, with a controlled representation of expressions through well-defined parameters. It is important to emphasize that the studies in Sign Languages must be constant since the language, as well as the technological resources, evolve dynamically. Therefore, it is expected that more research into BSL synthesis using virtual environments can be boosted in the coming years.

7.1 PUBLICATIONS AND ACKNOWLEDGMENT

This research had financial support from the Coordination of Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPQ), which supported directly or indirectly the progress of this research, being fundamental institutions in the technological and scientific development of the Brazil.

We thank the Department of Informatics of the Universidade Federal do Paraná (UFPR), where the research was developed. Thanks also to Friedrich Alexander Universitat for the support and co-advising in the first half of 2017.

The results of this work were published and released, up to now, in the course of their experiments in the following publications:

- GONÇALVES, D. A.; TODT, E. ; GARCIA, L. S.; "3D Avatar for automatic synthesis of signs for the sign languages." In: WSCG International Conferences in Central Europe on Computer Graphics, Visualization and Computer Vision, 2015, Plzen.

- GONÇALVES, D. A.; TODT, E.; "Extraction and Data Parsing of 4D Geometric Surfaces." In: V Symposium of Computational Numerical Methods, 2015, Curitiba. Computer Graphics.

- GONÇALVES, D. A.; TODT, E.; CLAUDIO, D. P.; "Finding Facial Expression's Principal Components to Support Speeding of a Sign Language Avatar Animation." In: XV Brazilian Symposium on Human Factors in Computational Systems, 2016, São Paulo. IHC16.

- GONÇALVES, D. A.; TODT, E.; CLÁUDIO, D. P. .; "Landmark-based Facial Expression Parametrization for Sign Languages Avatar Animation." In: Brazilian Symposium on

Human Factors in Computing Systems. New York: ACM Press, 2017. v. 16. p. 1. ISBN: 978-1-4503-6377-8, DOI:10.1145/3160504.3160507.

Other publications are under review without a final decision at this moment.

REFERENCES

Adhan, S. e Pintavirooj, C. (2016). Thai sign language recognition by using geometric invariant feature and ann classification. Em *2016 9th Biomedical Engineering International Conference (BMEiCON)*, páginas 1–4.

Agus, M., Marton, F., Bettio, F., Hadwiger, M. e Gobbetti, E. (2017). Data-driven analysis of virtual 3d exploration of a large sculpture collection in real-world museum exhibitions. *J. Comput. Cult. Herit.*, 11(1):2:1–2:20.

Ahire, A. L., Evans, A. e Blat, J. (2015). Animation on the web: A survey. Em *Proceedings of the 20th International Conference on 3D Web Technology*, Web3D '15, páginas 249–257, New York, NY, USA. ACM.

Aitpayev, K., Islam, S. e Imashev, A. (2016). Semi-automatic annotation tool for sign languages. Em *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, páginas 1–4.

Aldrich, J. (1998). *Doing Least Squares: Perspectives from Gauss and Yule*. International Statistical Review. 66 (1): 61–81.

Aleotti, J., Caselli, S. e Maccherozzi, G. (2005). Trajectory reconstruction with nurbs curves for robot programming by demonstration. Em *2005 International Symposium on Computational Intelligence in Robotics and Automation*, páginas 73–78.

Alkawaz, M. H. e Basori, A. H. (2012). The effect of emotional colour on creating realistic expression of avatar. Em *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '12, páginas 143–152, New York, NY, USA. ACM.

Amini, F., R. S. H. Z. V. Q. I. P. e McGuffin, M. J. (2014). The impact of interactivity on comprehending 2d and 3d visualizations of movement data. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):1–1.

Aouiti, N., Jemni, M. e Semreen, S. (2017). Arab gloss and implementation for arabic sign language. Em *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, páginas 1–6.

Azahar, M. A. B. M., Sunar, M. S. e Daman, D. (2010). *Low Computational Cost Crowd Rendering Method for Real-Time Virtual Heritage Environment*. electronic Journal of Computer Science and Information Technology (eJCSIT), Vol 2(12).

Balci, K. (2004). Xface: Mpeg-4 based open source toolkit for 3d facial animation. Em *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, páginas 399–402, New York, NY, USA. ACM.

Baoming, H. e Xuena, L. (2013). The rapid topological reconstruction of 3d-solid and simplification of weighted qem. Em *Vehicular Electronics and Safety (ICVES), 2013 IEEE International Conference on*, páginas 273–277.

Basawapatna, A., Repenning, A., Savignano, M., Manera, J., Escherle, N. e Repenning, L. (2018). Is drawing video game characters in an hour of code activity a waste of time? Em *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE 2018, páginas 93–98, New York, NY, USA. ACM.

Ben Yahia, N. e Jemni, M. (2013). A greedy inverse kinematics algorithm for animating 3d signing avatars. Em *Information and Communication Technology and Accessibility (ICTA), 2013 Fourth International Conference on*, páginas 1–3.

Bento, J., Claudio, A. e Urbano, P. (2014). Avatars on portuguese sign language. Em *Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on*, páginas 1–7.

Bouzid, Y., El Ghoul, O. e Jemni, M. (2013). Synthesizing facial expressions for signing avatars using mpeg4 feature points. Em *Information and Communication Technology and Accessibility (ICTA), 2013 Fourth International Conference on*, páginas 1–6.

Caesar, R., Suyoto e Gunanto, S. G. (2016). An automatic 3d face model segmentation for acquiring weight motion area. Em *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, páginas 81–86.

Cantwell, B., Warner, P., Koperwas, M. e Bhat, K. (2016). Ilm facial performance capture. Em *ACM SIGGRAPH 2016 Talks*, SIGGRAPH '16, páginas 26:1–26:2, New York, NY, USA. ACM.

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

Casas, D., Feng, A., Alexander, O., Fyffe, G., Debevec, P., Ichikari, R., Li, H., Olszewski, K., Suma, E. e Shapiro, A. (2016). Rapid photorealistic blendshape modeling from rgb-d sensors. Em *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, CASA '16, páginas 121–129, New York, NY, USA. ACM.

Cetinaslan, O. e Orvalho, V. (2018). Direct manipulation of blendshapes using a sketch-based interface. Em *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, páginas 14:1–14:10, New York, NY, USA. ACM.

Chu, A., Fu, C.-W., Hanson, A. e Heng, P.-A. (2009). Gl4d: A gpu-based architecture for interactive 4d visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1587–1594.

Connan, J. e Moemedi, K. A. (2010). Rendering an avatar from sign writing notation for sign language animation. Em *University of the Western Cape, Thesis*.

Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J. e Cottrell, G. W. (2010). *Evidence and a computational explanation of cultural differences in facial expression recognition*. Emotion, Vol 10(6).

Daoudi, M., Coello, Y., Descrosiers, P. A. e Ott, L. (2018). A new computational approach to identify human social intention in action. Em *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, páginas 512–516.

Debuchi, R. (2017). Bot3d editor, easy 3d computer character animation editor for smartphones and tablets. Em *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, SA '17, páginas 7:1–7:2, New York, NY, USA. ACM.

Dias, L., Carvalho, A. e Coelho, A. (2016). Spatial-temporal data watch to digital media: Phd thesis proposal in informatics engineering. Em *2016 11th Iberian Conference on Information Systems and Technologies (CISTI)*, páginas 1–5.

Draman, M. e Zeki, A. (2014). 3d animation as an alternative mean of da'wah for children. Em *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on*, páginas 1–5.

Elons, A., Ahmed, M. e Shedid, H. (2014). Facial expressions recognition for arabic sign language translation. Em *Computer Engineering Systems (ICCES), 2014 9th International Conference on*, páginas 330–335.

Erwig, M., Güting, R. H., Schneider, M. e Vazirgiannis, M. (1998). Abstract and discrete modeling of spatio-temporal data types. Em *Proceedings of the 6th ACM International Symposium on Advances in Geographic Information Systems*, GIS '98, páginas 131–136, New York, NY, USA. ACM.

Fang, T., Zhao, X., Shah, S. e Kakadiaris, I. (2011). 4d facial expression recognition. Em *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, páginas 1594–1601.

Felipe, T. A. (1997). *Introdução à gramática da LIBRAS*. Série Atualidades Pedagógicas 4 (3), 81-107.

Felipe, T. A. (2006). *Os processos de formação de palavra na Libras*. ETD-Educação Temática Digital 7 (2), 200-217.

Felipe, T. A. (2007). *Libras em contexto*. MS MONTEIRO, MEC/SEESP Nº Edição 7.

Feng, K. e Shen, Z. (2017). Compressing trajectory for trajectory indexing. Em *Proceedings of the 2Nd International Conference on Crowd Science and Engineering*, ICCSE'17, páginas 68–71, New York, NY, USA. ACM.

Feng, R. e Prabhakaran, B. (2016). On the "face of things". Em *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, páginas 3–4, New York, NY, USA. ACM.

Ferstl, Y., Kokkinara, E. e McDonnell, R. (2016). Do i trust you, abstract creature?: A study on personality perception of abstract virtual faces. Em *Proceedings of the ACM Symposium on Applied Perception*, SAP '16, páginas 39–43, New York, NY, USA. ACM.

Flotyński, J., Krzyszkowski, M. e Walczak, K. (2018). Query-based composition of animations for 3d web applications. Em *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, páginas 15:1–15:9, New York, NY, USA. ACM.

Flotyński, J. e Sobociński, P. (2018). Semantic 4-dimensionai modeling of vr content in a heterogeneous collaborative environment. Em *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, páginas 11:1–11:10, New York, NY, USA. ACM.

Gajalakshmi, P. e Sharmila, T. S. (2017). Sign language recognition of invariant features based on multiclass support vector machine with beam ecoc optimization. Em *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, páginas 587–591.

Gao, L., Lai, Y.-K., Liang, D., Chen, S.-Y. e Xia, S. (2016). Efficient and flexible deformation representation for data-driven surface modeling. *ACM Trans. Graph.*, 35(5):158:1–158:17.

Geng, L., Ma, X., Wang, H., Gu, J. e Li, Y. (2014). Chinese sign language recognition with 3d hand motion trajectories and depth images. Em *Proceeding of the 11th World Congress on Intelligent Control and Automation*, páginas 1457–1461.

Gloderer, M. e Hertle, A. (2010). Spline-based trajectory optimization for autonomous vehicles with ackerman drive.

Gonçalves, D. A. (2016). Voxelização de texturas na identificação de elementos dinâmicos em superfícies 3d. Em *in Proc. SMNE2016, Curitiba-Paraná, Brasil, pp.I:432-439.*

Gonçalves, D. A., Todt, E. e laura Sanchez Garcia (2015). 3d avatar for automatic synthesis of signs for the sign languages. Em *Proceedings, WSCG '2015, 23rd International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*.

Goran J. Zajić, Milena D. Vesić, A. M. G. e Reljin, I. S. (2016.). Animation content in frame analysis. Em *Telfor Journal, Vol. 8, No. 2.*, páginas 110–114.

Grif, M. e Manueva, Y. (2016). Semantic analyses of text to translate to russian sign language. Em *2016 11th International Forum on Strategic Technology (IFOST)*, páginas 286–289.

Guo, J. M., Tseng, S. H. e Wong, K. (2016). Accurate facial landmark extraction. *IEEE Signal Processing Letters*, 23(5):605–609.

Hada, M., Yamada, R. e Akamatsu, S. (2018). How does the transformation of an avatar face giving a favorable impression affect human recognition of the face? Em *2018 International Workshop on Advanced Image Technology (IWAIT)*, páginas 1–3.

Han, Y. (2015). 2d-to-3d visual human motion converting system for home optical motion capture tool and 3-d smart tv. *Systems Journal, IEEE*, 9(1):131–140.

Happy, S. e Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *Affective Computing, IEEE Transactions on*, 6(1):1–12.

Harold Whitaker, John Halas, T. S. (2009). *Timing for Animation*. Focal Press. Second Edition.

Hong, R., Yan, S. e Zhang, Z. (2015). Visual understanding with rgb-d sensors: An introduction to the special issue. *ACM Trans. Intell. Syst. Technol.*, 6(2):11:1–11:3.

Hongjian, W., Xuelian, Z., Hongli, L. e Xin, X. (2016). Spatial-temporal data modeling and visualizing method for uuv environmental perception. Em *2016 Chinese Control and Decision Conference (CCDC)*, páginas 5395–5402.

Hovden, G. e Ling, N. (2003). Optimizing facial animation parameters for mpeg-4. *IEEE Transactions on Consumer Electronics*, 49(4):1354–1359.

Huang, Y. (2017). Towards accurate marker-less human shape and pose estimation over time. Em *2017 International Conference on 3D Vision (3DV)*, páginas 421–430.

Huang, Y. e Khan, S. (2016). Mirroring facial expressions: Evidence from visual analysis of dyadic interactions. Em *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, páginas 225–228, New York, NY, USA. ACM.

Huenerfauth, M. (2016). Interspeech - selecting exemplar recordings of american sign language non-manual expressions for animation synthesis based on manual sign timing. Em *Speech and Language Processing for Assistive Technologies. SLPAT*.

Huenerfauth, M., Lu, P. e Rosenberg, A. (2011). Evaluating importance of facial expression in american sign language and pidgin signed english animations. Em *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, páginas 99–106, New York, NY, USA. ACM.

Hyde, J., Carter, E. J., Kiesler, S. e Hodgins, J. K. (2016). Evaluating animated characters: Facial motion magnitude influences personality perceptions. *ACM Trans. Appl. Percept.*, 13(2):8:1–8:17.

Iatskiu, C. E. A., García, L. S. e Antunes, D. R. (2017). Automatic signwriting generation of libras signs from core-sl. Em *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, IHC 2017, páginas 55:1–55:4, New York, NY, USA. ACM.

Idaka, Y., Yasuda, K., Ho, Y. e Tagawa, N. (2017). Cost-effective camera pose estimation for basketball analysis using radon transform. Em *2017 6th International Conference on Informatics, Electronics and Vision 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, páginas 1–6.

Jain, S., Ye, Y. e Liu, C. K. (2009). Optimization-based interactive motion synthesis. *ACM Trans. Graph.*, 28(1):10:1–10:12.

Jan, A. e Meng, H. (2015). Automatic 3d facial expression recognition using geometric and textured feature fusion. Em *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 05, páginas 1–6.

Ji, Y., Liu, C., Gong, S. e Cheng, W. (2016). 3d hand gesture coding for sign language learning. Em *2016 International Conference on Virtual Reality and Visualization (ICVRV)*, páginas 407–410.

Jian, W., Hai-Ling, W., Bo, Z. e Ni, J. (2013). An efficient mesh simplification method in 3d graphic model rendering. Em *Internet Computing for Engineering and Science (ICICSE), 2013 Seventh International Conference on*, páginas 55–59.

Johnson, E. R. e Murphey, T. D. (2009). Automated trajectory synthesis from animation data using trajectory optimization. Em *2009 IEEE International Conference on Automation Science and Engineering*, páginas 274–279.

Juanes-Méndez, J. A., Palomera, P. R., Briz-Ponce, L. e Ledesma, M. J. S. (2016). 4d visual environment on mobile devices for learning in the human anatomy field. Em *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM '16, páginas 467–471, New York, NY, USA. ACM.

Kacorri, H. (2013). Models of linguistic facial expressions for american sign language animation. *SIGACCESS Access. Comput.*, (105):19–23.

Kacorri, H. (2015). *TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation*. CUNY Academic Works.

Kacorri, H. e Huenerfauth, M. (2014). Implementation and evaluation of animation controls sufficient for conveying asl facial expressions. Em *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, páginas 261–262, New York, NY, USA. ACM.

Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K. e Willard, M. (2015). Demographic and experiential factors influencing acceptance of sign language animation by deaf users. Em *Proceedings of the 17th International ACM SIGACCESS Conference on Computers &#38; Accessibility*, ASSETS '15, páginas 147–154, New York, NY, USA. ACM.

Kai, Z., Feng, I., Guo, A. e Zhong, E. (2014). A clustering compression method for 3d human motion capture data. Em *Computer Science Education (ICCSE), 2014 9th International Conference on*, páginas 781–784.

Kaur, S. e Singh, M. (2015). Indian sign language animation generation system. Em *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*, páginas 909–914.

Ko, M.-C. e Choy, Y.-C. (2002). 3d mesh simplification for effective network transmission. Em *High Speed Networks and Multimedia Communications 5th IEEE International Conference on*, páginas 284–288.

Kocon, M. (2014). Facial expressions modeling for interactive virtual environments. Em *Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On*, páginas 744–747.

Krispel, U., Settgast, V. e Fellner, D. W. (2018). Dynamo - dynamic 3d models for the web: A declarative approach to dynamic and interactive 3d models on the web using x3dom. Em *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, páginas 16:1–16:5, New York, NY, USA. ACM.

Laptev, I. e Lindeberg, T. (2003). Space-time interest points. Em *in Proc. ICCV'03, Nice, France, pp.I:432-439.*

Le, V., Tang, H. e Huang, T. (2011). Expression recognition from 3d dynamic faces using robust spatio-temporal shape features. Em *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, páginas 414–421.

Lee, J., Han, B. e Choi, S. (2016). Interactive motion effects design for a moving object in 4d films. Em *Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology*, VRST '16, páginas 219–228, New York, NY, USA. ACM.

Lee, Y., Lee, S. J. e Popović, Z. (2009). Compact character controllers. *ACM Trans. Graph.*, 28(5):169:1–169:8.

Lemaire, P., Ben Amor, B., Ardabilian, M., Chen, L. e Daoudi, M. (2011). Fully automatic 3d facial expression recognition using a region-based approach. Em *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, páginas 53–58, New York, NY, USA. ACM.

Li, H., Kulik, L. e Ramamohanarao, K. (2014). Spatio-temporal trajectory simplification for inferring travel paths. Em *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, páginas 63–72, New York, NY, USA. ACM.

Li, N., Rea, D. J., Young, J. E., Sharlin, E. e Sousa, M. C. (2015). And he built a crooked camera: A mobile visualization tool to view four-dimensional geometric objects. Em *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, SA '15, páginas 23:1–23:5, New York, NY, USA. ACM.

Li, X., Sheng, B., Li, P., Kim, J. e Feng, D. D. (2018). Voxelized facial reconstruction using deep neural network. Em *Proceedings of Computer Graphics International 2018*, CGI 2018, páginas 1–4, New York, NY, USA. ACM.

Li, X., Yu, J., Gao, F. e Zhang, J. (2016). Data-driven facial animation via hypergraph learning. Em *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, páginas 000442–000445.

Lombardo, V., Battaglino, C., Damiano, R. e Nunnari, F. (2011). An avatar-based interface for the italian sign language. Em *Complex, Intelligent and Software Intensive Systems (CISIS), 2011 International Conference on*, páginas 589–594.

Lundqvist, D., . L. J. E. (1998). The averaged karolinska directed emotional faces - akdef. Em *CD ROM from Department of Clinical Neuroscience, Psychology section*.

Lv, S., Da, F. e Deng, X. (2015). A 3d face recognition method using region-based extended local binary pattern. Em *Image Processing (ICIP), 2015 IEEE International Conference on*, páginas 3635–3639.

Lyons, M. J., Akemastu, S., Kamachi, M. e Gyoba, J. (1998). *Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition*.

Mahmoud, M. M., Baltrušaitis, T. e Robinson, P. (2014). Automatic detection of naturalistic hand-over-face gesture descriptors. Em *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, páginas 319–326, New York, NY, USA. ACM.

Maniadakis, M., Droit-Volet, S. e Trahanias, P. (2017). Emotionally modulated time perception for prioritized robot assistance. Em *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, páginas 197–198, New York, NY, USA. ACM.

Martin Erwig, M. S. e Güting, R. H. (1998). Temporal objects for spatio-temporal data models and a comparison of their representations. Em *Int. Workshop on Advances in Database Technologies*, LNCS 1552, páginas 454–465.

McGowen, V. e Geigel, J. (2016). Automatic blend shape creation for facial motion capture. Em *ACM SIGGRAPH 2016 Posters*, SIGGRAPH '16, páginas 44:1–44:2, New York, NY, USA. ACM.

Methirumangalath, S., Dev Parakkat, A., Kannan, S. S. e Muthuganapathy, R. (2017). Reconstruction using a simple triangle removal approach. Em *SIGGRAPH Asia 2017 Technical Briefs*, SA '17, páginas 27:1–27:4, New York, NY, USA. ACM.

Mkrtchyan, K., Chakraborty, A. e Roy-Chowdhury, A. (2016). Optimal landmark selection for registration of 4d confocal image stacks in arabidopsis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99):1–1.

Mori, H., Nakadai, T., Toyama, F. e Shoji, K. (2015). Gaze animation optimization based on a viewer's preference. Em *SIGGRAPH Asia 2015 Posters*, SA '15, páginas 14:1–14:1, New York, NY, USA. ACM.

Mueen, A. e Keogh, E. (2016). Extracting optimal performance from dynamic time warping. Em *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, páginas 2129–2130, New York, NY, USA. ACM.

Neidle, C., Bahan, B., MacLaughlin, D., Lee, R. G. e Kegl, J. (1998). Realizations of syntactic agreement in american sign language: Similarities between the clause and the noun phrase. *Studia Linguistica*, 52(3):191–226.

Ng, K.-W. e Low, Z.-W. (2014). Simplification of 3d triangular mesh for level of detail computation. Em *Computer Graphics, Imaging and Visualization (CGIV), 2014 11th International Conference on*, páginas 11–16.

Ngueyep, R, S. N. (2014). Large vector auto regressions for multi-layer spatially correlated time series. Em *Technometrics, in press with online version available*.

Obaid, M., Mukundan, R., Billinghurst, M. e Pelachaud, C. (2010). Expressive mpeg-4 facial animation using quadratic deformation models. Em *Computer Graphics, Imaging and Visualization (CGIV), 2010 Seventh International Conference on*, páginas 9–14.

Oh, E., Lee, M. e Lee, S. (2011). How 4d effects cause different types of presence experience? Em *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*, VRCAI '11, páginas 375–378, New York, NY, USA. ACM.

Oliveira, M., Chatbri, H., Little, S., O'Connor, N. E. e Sutherland, A. (2017). A comparison between end-to-end approaches and feature extraction based approaches for sign language recognition. Em *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, páginas 1–6.

Othman, A. e Jemni, M. (2017). An xml-gloss annotation system for sign language processing. Em *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, páginas 1–7.

Pan, Z., Di, M., Zhang, J. e Ravi, S. (2018). Automatic re-topology and uv remapping for 3d scanned objects based on neural network. Em *Proceedings of the 31st International Conference on Computer Animation and Social Agents*, CASA 2018, páginas 48–52, New York, NY, USA. ACM.

Park, I. K., Lee, S. W. e Lee, S. U. (2002). Shape-adaptive 3d mesh simplification based on local optimality measurement. Em *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, páginas 462–466.

Polys, N., Newcomb, C., Schenk, T., Skuzinski, T. e Dunay, D. (2018). The value of 3d models and immersive technology in planning urban density. Em *Proceedings of the 23rd International ACM Conference on 3D Web Technology*, Web3D '18, páginas 13:1–13:4, New York, NY, USA. ACM.

Punchimudiyanse, M. e Meegama, R. (2015). 3d signing avatar for sinhala sign language. Em *Industrial and Information Systems (ICIIS), 2015 IEEE 10th International Conference on*, páginas 290–295.

Ratan, R. e Hasler, B. S. (2014). Playing well with virtual classmates: Relating avatar design to group satisfaction. Em *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '14, páginas 564–573, New York, NY, USA. ACM.

Rieger, T. (2003). Avatar gestures. Em *WSCG 2003, International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, volume 2, páginas 379–386.

Sagar, M. A. (2016). Making simulated faces come alive. Em *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, páginas 1–1, Piscataway, NJ, USA. IEEE Press.

Sahid, A. F. B. M., Ismail, W. S. W. e Ghani, D. A. (2016). Malay sign language (msl) for beginner using android application. Em *2016 International Conference on Information and Communication Technology (ICICTM)*, páginas 189–193.

Sandbach, G., Zafeiriou, S., Pantic, M. e Rueckert, D. (2011). A dynamic approach to the recognition of 3d facial expressions and their temporal models. Em *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, páginas 406–413.

Schoch, M., Alonso-Mora, J., Siegwart, R. e Beardsley, P. (2014). Viewpoint and trajectory optimization for animation display with aerial vehicles. Em *2014 IEEE International Conference on Robotics and Automation (ICRA)*, páginas 4711–4716.

Serban, N. (2011). Space-time varying coeficient model: The equity of service distribution. Em *Annals of Applied Statistics 5*.

Setty, S. e Mudenagudi, U. (2018). Region of interest-based 3d inpainting of cultural heritage artifacts. *J. Comput. Cult. Herit.*, 11(2):9:1–9:21.

Shang, Z., Joshi, J. e Hoey, J. (2017). Continuous facial expression recognition for affective interaction with virtual avatar. Em *2017 IEEE International Conference on Image Processing (ICIP)*, páginas 1995–1999.

Sheidin, J., Lanir, J., Kuflik, T. e Bak, P. (2017). Visualizing spatial-temporal evaluation of news stories. Em *Proceedings of the 22Nd International Conference on Intelligent User Interfaces Companion*, IUI '17 Companion, páginas 65–68, New York, NY, USA. ACM.

Sikdar, B. (2017). Spatio-temporal correlations in cyber-physical systems: A defense against data availability attacks. Em *Proceedings of the 3rd ACM Workshop on Cyber-Physical System Security*, CPSS '17, páginas 103–110, New York, NY, USA. ACM.

Sofiato, Cássia Geciauskas, . R. L. H. (2014). Brazilian sign language dictionaries: comparative iconographical and lexical study. Em *Educação e Pesquisa, 40(1) 109-126. https://dx.doi.org/10.1590/S1517-97022014000100008*.

Soga, A., Yazaki, Y., Umino, B. e Hirayama, M. (2016). Body-part motion synthesis system for contemporary dance creation. Em *ACM SIGGRAPH 2016 Posters*, SIGGRAPH '16, páginas 29:1–29:2, New York, NY, USA. ACM.

Streuber, S., Quiros-Ramirez, M. A., Hill, M. Q., Hahn, C. A., Zuffi, S., O'Toole, A. e Black, M. J. (2016). Body talk: Crowdshaping realistic 3d avatars with words. *ACM Trans. Graph.*, 35(4):54:1–54:14.

Su, H., Zheng, K., Huang, J., Wang, H. e Zhou, X. (2015). Calibrating trajectory data for spatio-temporal similarity analysis. *The VLDB Journal*, 24(1):93–116.

Suheryadi, A. e Nugroho, H. (2016). Spatio-temporal analysis for moving object detection under complex environment. Em *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, páginas 498–505.

Sulfayanti, Dewiani e Lawi, A. (2016). A real time alphabets sign language recognition system using hands tracking. Em *2016 International Conference on Computational Intelligence and Cybernetics*, páginas 69–72.

Sun, Y. e Yin, L. (2008). Facial expression recognition based on 3d dynamic range model sequences. Em *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, páginas 58–71, Berlin, Heidelberg. Springer-Verlag.

Szwoch, W. (2015). Model of emotions for game players. Em *Human System Interactions (HSI), 2015 8th International Conference on*, páginas 285–290.

T. Jusko, E. S. (2016). Scalable trajectory optimization based on bézier curves.

Talbi, I., Ghoul, O. E. e Jemni, M. (2017). Towards realistic simulation of facial deformation in sign language. Em *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, páginas 1–5.

Tanaka, H., Sakriani, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H. e Nakamura, S. (2016). Teaching social communication skills through human-agent interaction. *ACM Trans. Interact. Intell. Syst.*, 6(2):18:1–18:26.

Tangsuksant, W., Adhan, S. e Pintavirooj, C. (2014). American sign language recognition by using 3d geometric invariant feature and ann classification. Em *Biomedical Engineering International Conference (BMEiCON), 2014 7th*, páginas 1–5.

Tim, S., Rombaut, M. e Pellerin, D. (2014). Dictionary of gray-level 3d patches for action recognition. Em *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, páginas 1–6.

Troy, Pranowo e Gunanto, S. G. (2016). 2d to 3d space transformation for facial animation based on marker data. Em *2016 6th International Annual Engineering Seminar (InAES)*, páginas 1–5.

Vemulapalli, R., Arrate, F. e Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. Em *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, páginas 588–595.

Vu, H., Nguyen, V.-G., Pham, A.-T. e Tran, T.-H. (2017). Pedestrian localization and trajectory reconstruction in a surveillance camera network. Em *Proceedings of the Eighth International Symposium on Information and Communication Technology*, SoICT 2017, páginas 393–400, New York, NY, USA. ACM.

Wantroba, E. J. e Romero, R. A. F. (2015). An interactive question-answer system with dialogue for a receptionist avatar. Em *2015 12th Latin American Robotics Symposium and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR)*, páginas 360–365.

Wauck, H., Lucas, G., Shapiro, A., Feng, A., Boberg, J. e Gratch, J. (2018). Analyzing the effect of avatar self-similarity on men and women in a search and rescue game. Em *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, páginas 485:1–485:12, New York, NY, USA. ACM.

Wiegand, K. (2014). Intelligent assistive communication and the web as a social medium. Em *Proceedings of the 11th Web for All Conference*, W4A '14, páginas 27:1–27:2, New York, NY, USA. ACM.

Wu, Y. (2016). Facial landmark detection and tracking for facial behavior analysis. Em *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, páginas 431–434, New York, NY, USA. ACM.

Xu, W., Miao, Z., Zhang, X. P. e Tian, Y. (2017). A hierarchical spatio-temporal model for human activity recognition. *IEEE Transactions on Multimedia*, 19(7):1494–1509.

Xue, M., Mian, A., Liu, W. e Li, L. (2014). Fully automatic 3d facial expression recognition using local depth features. Em *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, páginas 1096–1103.

Xue, M., Mian, A., Liu, W. e Li, L. (2015). Automatic 4d facial expression recognition using dct features. Em *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, páginas 199–206.

Yamane, K. e Goerner, J. (2014). Task assignment and trajectory optimization for displaying stick figure animations with multiple mobile robots. Em *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, páginas 3806–3813.

Yamina, B. e Farida, M. (2012). Mpeg4 parameterization for facial deformation. Em *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*, páginas 414–419.

Yan, P., Khan, S. e Shah, M. (2008). Learning 4d action feature models for arbitrary view action recognition. Em *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, páginas 1–7.

Yao, Y., Huang, D., Yang, X., Wang, Y. e Chen, L. (2018). Texture and geometry scattering representation-based facial expression recognition in 2d+3d videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 14(1s):18:1–18:23.

Yashar Taghizadegan, Hassan Ghassemian, M. N.-M. (2012). 3d face recognition method using 2dpca-euclidean distance classification. *ACEEE International Journal on Control System and Instrumentation*, 3(1):5.

Yu, S. e Poger, S. (2017). Using a temporal weighted data model to maximize influence in mobile messaging apps for computer science education. *J. Comput. Sci. Coll.*, 32(6):210–211.

Zhai, G., Ren, F., Zhang, G. e Evison, M. (2011). Facial shape analysis based on euclidean distance matrix analysis. Em *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, volume 4, páginas 1896–1900.

Zhang, H., Weng, J. e Ruan, G. (2014). Visualizing 2-dimensional manifolds with curve handles in 4d. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2575–2584.

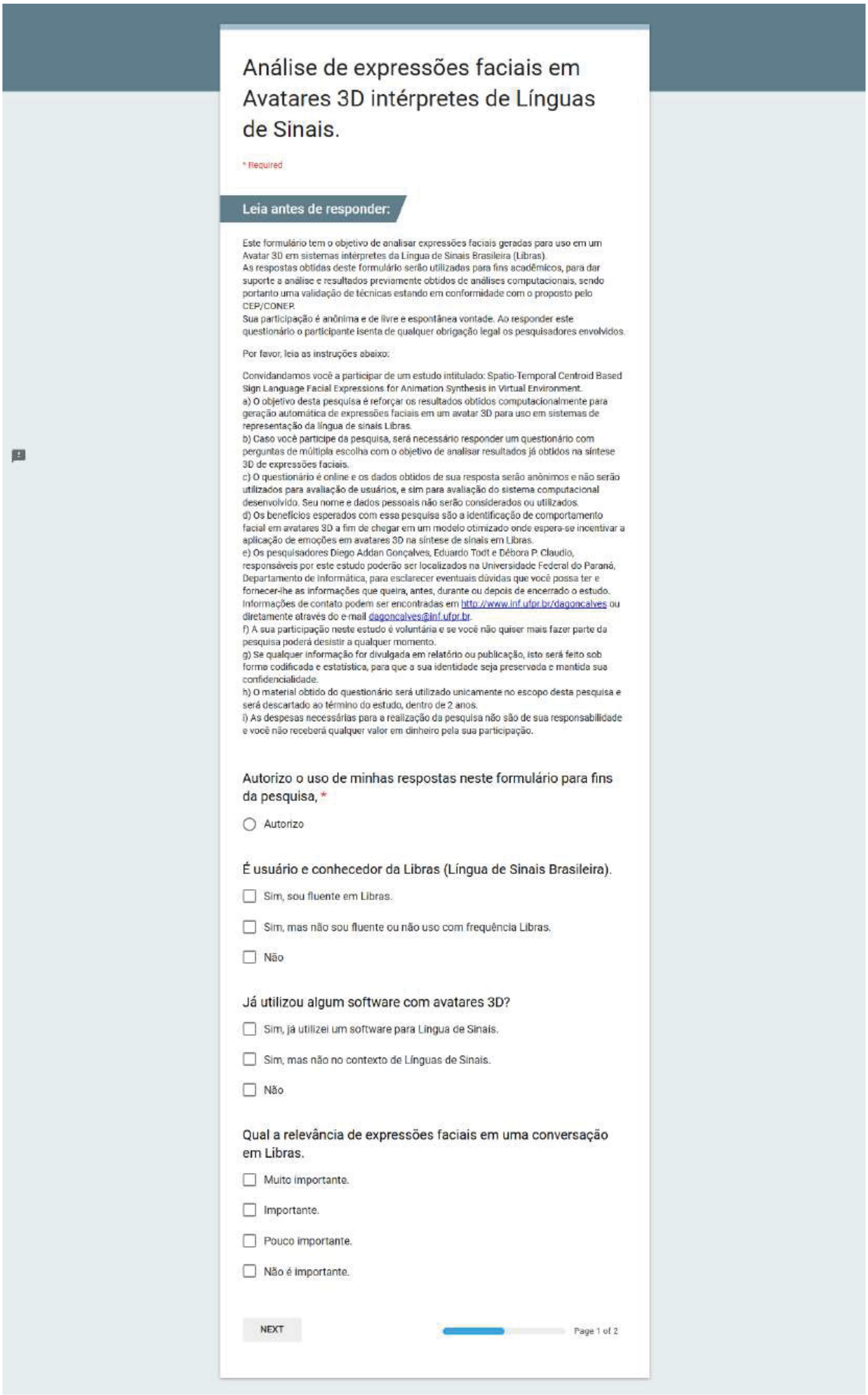
Zhao, M. e Zhang, J. (2008). Rapidly product and optimize facial animation methods for 3d game. Em *2008 International Conference on Internet Computing in Science and Engineering*, páginas 136–139.

Zhao, Y., Jiang, D. e Sahli, H. (2015). 3d emotional facial animation synthesis with factored conditional restricted boltzmann machines. Em *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, páginas 797–803.

Zhu, Y., Chen, W. e Guo, G. (2015). Fusing multiple features for depth-based action recognition. *ACM Trans. Intell. Syst. Technol.*, 6(2):18:1–18:20.

Zou, C., Chen, S., Fu, H. e Liu, J. (2015). Progressive 3d reconstruction of planar-faced manifold objects with drf-based line drawing decomposition. *Visualization and Computer Graphics, IEEE Transactions on*, 21(2):252–263.

APPENDIX A – DATA-SET OUTPUTS FORM



Análise de expressões faciais em Avatares 3D intérpretes de Línguas de Sinais.

* Required

Leia antes de responder:

Este formulário tem o objetivo de analisar expressões faciais geradas para uso em um Avatar 3D em sistemas intérpretes da Língua de Sinais Brasileira (Libras).
As respostas obtidas deste formulário serão utilizadas para fins acadêmicos, para dar suporte a análise e resultados previamente obtidos de análises computacionais, sendo portanto uma validação de técnicas estando em conformidade com o proposto pelo CEP/CONEP.
Sua participação é anônima e de livre e espontânea vontade. Ao responder este questionário o participante isenta de qualquer obrigação legal os pesquisadores envolvidos.

Por favor, leia as instruções abaixo:

Convidandamos você a participar de um estudo intitulado: Spatio-Temporal Centroid Based Sign Language Facial Expressions for Animation Synthesis in Virtual Environment.
a) O objetivo desta pesquisa é reforçar os resultados obtidos computacionalmente para geração automática de expressões faciais em um avatar 3D para uso em sistemas de representação da língua de sinais Libras.
b) Caso você participe da pesquisa, será necessário responder um questionário com perguntas de múltipla escolha com o objetivo de analisar resultados já obtidos na síntese 3D de expressões faciais.
c) O questionário é online e os dados obtidos de sua resposta serão anônimos e não serão utilizados para avaliação de usuários, e sim para avaliação do sistema computacional desenvolvido. Seu nome e dados pessoais não serão considerados ou utilizados.
d) Os benefícios esperados com essa pesquisa são a identificação de comportamento facial em avatares 3D a fim de chegar em um modelo otimizado onde espera-se incentivar a aplicação de emoções em avatares 3D na síntese de sinais em Libras.
e) Os pesquisadores Diego Addan Gonçalves, Eduardo Todt e Débora P. Claudio, responsáveis por este estudo poderão ser localizados na Universidade Federal do Paraná, Departamento de Informática, para esclarecer eventuais dúvidas que você possa ter e fornecer-lhe as informações que queira, antes, durante ou depois de encerrado o estudo. Informações de contato podem ser encontradas em http://www.inf.ufpr.br/dagoncalves ou diretamente através do e-mail dagoncalves@inf.ufpr.br.
f) A sua participação neste estudo é voluntária e se você não quiser mais fazer parte da pesquisa poderá desistir a qualquer momento.
g) Se qualquer informação for divulgada em relatório ou publicação, isto será feito sob forma codificada e estatística, para que a sua identidade seja preservada e mantida sua confidencialidade.
h) O material obtido do questionário será utilizado unicamente no escopo desta pesquisa e será descartado ao término do estudo, dentro de 2 anos.
i) As despesas necessárias para a realização da pesquisa não são de sua responsabilidade e você não receberá qualquer valor em dinheiro pela sua participação.

Autorizo o uso de minhas respostas neste formulário para fins da pesquisa, *

○ Autorizo

É usuário e conhecedor da Libras (Língua de Sinais Brasileira).

□ Sim, sou fluente em Libras.

□ Sim, mas não sou fluente ou não uso com frequência Libras.

□ Não

Já utilizou algum software com avatares 3D?

□ Sim, já utilizei um software para Língua de Sinais.

□ Sim, mas não no contexto de Línguas de Sinais.

□ Não

Qual a relevância de expressões faciais em uma conversação em Libras.

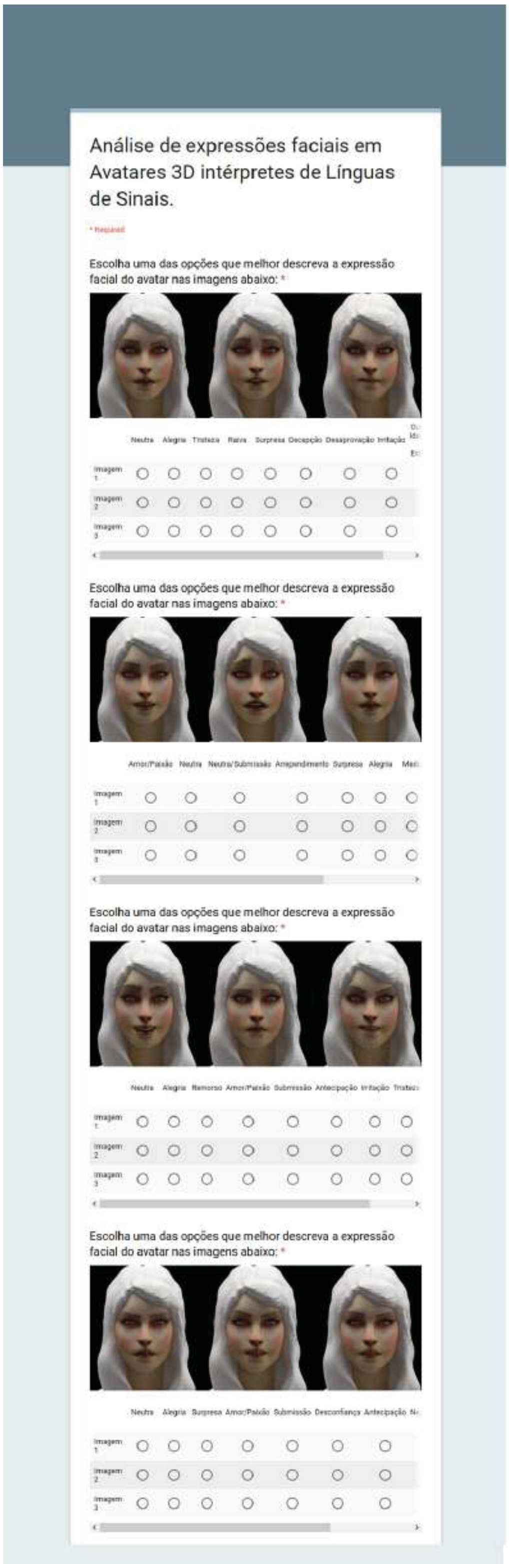
□ Muito importante.

□ Importante.

□ Pouco importante.

□ Não é importante.

Figure A.1: Page 1 of the online form. Here basic information was required as the user's familiarity with the BSL language, as well as legal information about the research.



Análise de expressões faciais em Avatares 3D intérpretes de Línguas de Sinais.

* Required

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Alegria	Tristeza	Raiva	Surpresa	Decepção	Desaprovação	Irritação	D Id Ei
Imagem 1	○	○	○	○	○	○	○	○	
Imagem 2	○	○	○	○	○	○	○	○	
Imagem 3	○	○	○	○	○	○	○	○	

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

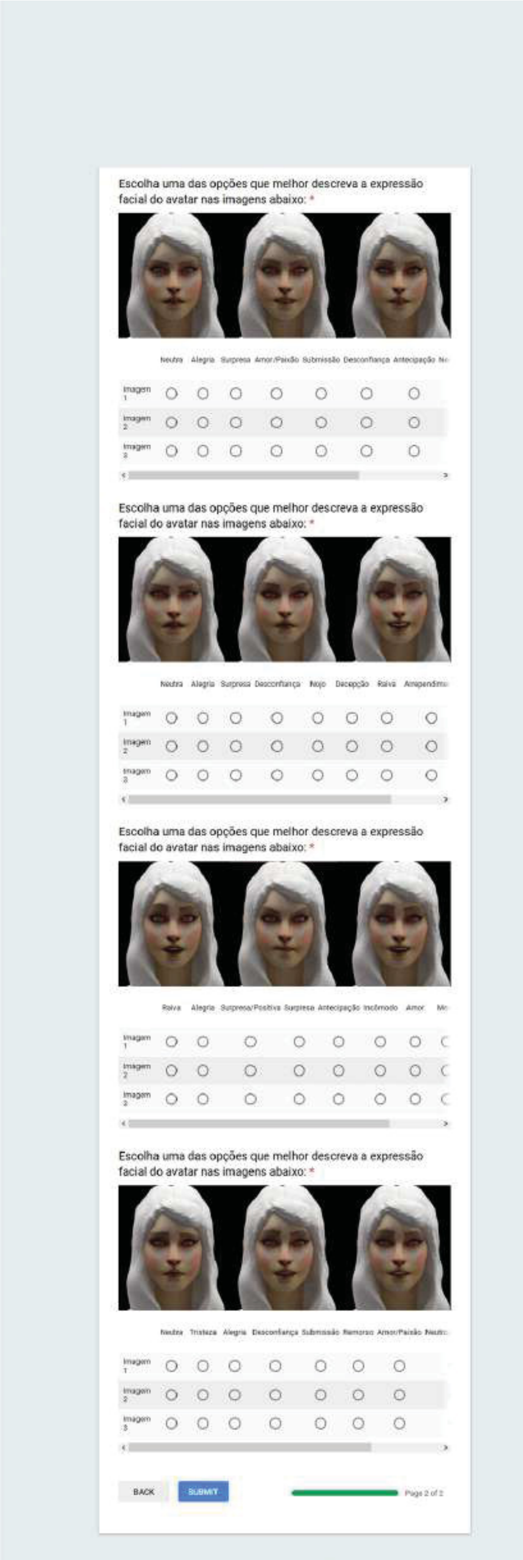
	Amor/Paixão	Neutra	Neutra/Submissão	Arrependimento	Surpresa	Alegria	Me
Imagem 1	○	○	○	○	○	○	○
Imagem 2	○	○	○	○	○	○	○
Imagem 3	○	○	○	○	○	○	○

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Alegria	Remorso	Amor/Paixão	Submissão	Antecipação	Irritação	Tristez
Imagem 1	○	○	○	○	○	○	○	○
Imagem 2	○	○	○	○	○	○	○	○
Imagem 3	○	○	○	○	○	○	○	○

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Alegria	Surpresa	Amor/Paixão	Submissão	Desconfiança	Antecipação	N
Imagem 1	○	○	○	○	○	○	○	
Imagem 2	○	○	○	○	○	○	○	
Imagem 3	○	○	○	○	○	○	○	



Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Alegria	Surpresa	Amor/Paixão	Submissão	Desconfiança	Antecipação	N
Imagem 1	○	○	○	○	○	○	○	
Imagem 2	○	○	○	○	○	○	○	
Imagem 3	○	○	○	○	○	○	○	

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Alegria	Surpresa	Desconfiança	Nojo	Decepção	Raiva	Arrependim
Imagem 1	○	○	○	○	○	○	○	○
Imagem 2	○	○	○	○	○	○	○	○
Imagem 3	○	○	○	○	○	○	○	○

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Raiva	Alegria	Surpresa/Positiva	Surpresa	Antecipação	Incômodo	Amor	M
Imagem 1	○	○	○	○	○	○	○	○
Imagem 2	○	○	○	○	○	○	○	○
Imagem 3	○	○	○	○	○	○	○	○

Escolha uma das opções que melhor descreva a expressão facial do avatar nas imagens abaixo: *

	Neutra	Tristeza	Alegria	Desconfiança	Submissão	Remorso	Amor/Paixão	Neutr
Imagem 1	○	○	○	○	○	○	○	
Imagem 2	○	○	○	○	○	○	○	
Imagem 3	○	○	○	○	○	○	○	

Figure A.2: Page 2 of the online form. Here we present exits of base expressions, interpolations and simplifications obtained from the constructed dataset.