# Facial Expressions Animation in Sign Language based on Spatio-temporal Centroid

Diego Addan Gonçalves[1], Maria Cecília Calani Baranauskas[1], Julio César dos Reis[1] and Eduardo Todt[2]

[1]*Institute of Computing, University of Campinas, São Paulo, Brazil*

[2]*Department of Informatics, Universidade Federal do Paraná, Curitiba, Brazil*

Keywords: 3D Avatar, Sign Language, Facial Expression.

Abstract: Systems that use virtual environments with avatars for information communication are of fundamental importance in contemporary life. They are even more relevant in the context of supporting sign language communication for accessibility purposes. Although facial expressions provide message context and define part of the information transmitted, *e.g.*, irony or sarcasm, facial expressions are usually considered as a static background feature in a primarily gestural language in computational systems. This article proposes a novel parametric model of facial expression synthesis through a 3D avatar representing complex facial expressions leveraging emotion context. Our technique explores interpolation of the base expressions in the geometric animation through centroids control and Spatio-temporal data. The proposed method automatically generates complex facial expressions with controllers that use region parameterization as in manual models used for sign language representation. Our approach to the generation of facial expressions adds emotion to representation, which is a determining factor in defining the tone of a message. This work contributes with the definition of non-manual markers for Sign Languages 3D Avatar and the refinement of the synthesized message in sign languages, proposing a complete model for facial parameters and synthesis using geometric centroid regions interpolation. A dataset with facial expressions was generated using the proposed model and validated using machine learning algorithms. In addition, evaluations conducted with the deaf community showed a positive acceptance of the facial expressions and synthesized emotions.

## 1 INTRODUCTION

Systems that use virtual environments for information communication are of fundamental importance in everyday life, providing flexibility and speed in the transmission of information (Lombardo et al., 2011) (Punchimudiyanse and Meegama, 2015). Studies on gesture synthesis in 3D space have an immediate impact on the development of new essential technologies and deserve attention and investment. 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) (Kacorri et al., 2015) (Sofiato, 2014), most of which based on recent studies.

Several computational systems have been developed for Sign Languages based on signal synthesis through the user interaction (Bento et al., 2014) (Adhan and Pintavirooj, 2016) (Ratan and Hasler, 2014), (Kacorri et al., 2015). In general, those systems use configuration parameters for hand-based gestures, body and arms positioning, aiming at the fidelity between the virtual and real representations. Although facial expressions provide message context and define part of the information transmitted (Elons et al., 2014), *e.g.*, irony or sarcasm, there is a serious lack of such features in most of the existing software tools. 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).

According to Neidle *et al.* (Neidle et al., 1998), it is remarkable 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. In this sense, the accurate transmission of a message using sign languages needs a facial expression representation as a feeling modifier or as a context supplement for the



Gonçalves, D., Baranauskas, M., Reis, J. and Todt, E.
Facial Expressions Animation in Sign Language based on Spatio-temporal Centroid.
DOI: 10.5220/0009344404630475
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 2*, pages 463-475
ISBN: 978-989-758-423-7

raw gesture information as well as for morphological signs.

However, facial parameters for the computational representation of sign languages are scarce and lack details that correspond with the models applied to the manual elements of the language. Facial expressions, that also provides the emotion, should consider cultural elements, morpho-syntactic elements and facial expression interpolations. Although some systems implement facial expressions in their avatars, they often do not follow any models integrated with the other parameters and do not associate the emotional expression with the message. To this end, it is necessary a control model that understands the relationship of each facial region with facial expressions for the generation of a virtual environment. A model with these characteristics facilitates the integration of complex features in the system.

The manual sign language parameters has well-defined models, such as the CORE-SL (Iatskiu et al., 2017), that describes the hand parameters as shape, location, movement and orientation. Each parameters have values and hierarchy relationship that compounds a signal. These models do not incorporate detailed facial parameters, or Non-Manual Markers (NMM), a fact that unfortunately holds for models and systems for other sign languages as well.

In this paper, we propose and validate a novel parametric model of facial expression applied to a 3D avatar dedicated to Sign Language synthesis. The animation of facial expressions can be controlled by understanding the behavior of facial landmarks and the geometric animations allowing the implementation of automatic signal synthesis and a facial model for sign languages. In particular, we propose a functional facial model for Sign Languages 3D animation that considers parameters of the face using Spatio-temporal information. The Spatio-temporal information enables to control the behavior of the geometric landmarks and to generate an automatic synthesis of emotions through a 3D avatar.

Our investigation involved the definition and application of a parameterized landmark-based model of facial expressions in a signal language synthesis system. We aimed to enhance the animation process supporting a more accurate representation of sign language messages and integrating into the Core-SL hierarchic model.

The conducted methodology is synthesized as follows:

- Proposed a parameterized computational model for facial representation in sign language 3D avatar.
- Identified the main components of deformation landmarks in the avatar face to define the relationship between facial regions and base expressions. Base expressions are the main emotions used to generate all the secondary interpolations. This enables to generate expressions that follow the descriptive sign language model by associating the expression with the signal message. This allows full control of the geometric mesh of the face.
- Applied concepts of temporal data related to a geometric mesh to control and optimize the 4D actions of a 3D avatar.
- Evaluated the proposed system through a data set based on interpolation of parameters and generation of simplified and adapted expressions. The evaluation was carried out in two steps: Using machine learning to classify avatar generated expressions and based on the feedback of the Deaf community concerning the outputs.

Our contribution offers a fine control of the geometric deformations in the representation of emotions during the transmission of the message, defining which parameters are the most relevant in the representation of the main expressions, allowing local control for each parameter. The main contributions are the definition of a NMM model for sign languages avatar and a process to synthesize emotions and expressions in a 3D avatar for synthesis systems, supported by the extraction of the spatio-temporal data of facial animation that defines facial behavior.

The remaining of this paper is organized as follows: Section 2 presents a synthesis of related work. Section 3 presents the built model for facial parameters in the 3D avatar. Section 4 reports on our technique for the automatic generation of complex facial expressions based on behaviors of facial landmarks. Section 5 describes the evaluations conducted to assess our proposal. Section 6 refers to the final remarks.

## 2 RELATED WORK

The use of virtual agents for educational or entertainment systems has increased as well as the interest by target users (Grif and Manueva, 2016) (Ratan and Hasler, 2014) (Wiegand, 2014). Basawapatna et al. (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.

The avatar motions may be generated by controllers associated with the 3D geometric mesh using techniques such as Blend Shapes or Morph Tar-

gets as well as the tracking of geometric controllers based on a set of features and classifier cascade (Feng and 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 (Kaur and Singh, 2015). These models make a parallel between sign language representation and their descriptive notations, but do not cover complex elements for facial definitions (Lombardo et al., 2011).

Modulation in the mouth, eyebrows and other points of interest in the face can change the meaning of expressions in sign languages (Huenerfauth et al., 2011), reinforcing the necessity of a facial model that defines complex elements of the face. There are facial expression definitions for sign languages such as the work of Elons which limits the categories of: *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 has paused the message momentarily.

These categories are still very generic and an ideal definition of facial parameters is similar to used for manual elements in the descriptive models of sign languages (Iatskiu et al., 2017) identifying independent elements and position values. 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 and Basori, 2012). 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.

The face can be decomposed into specific and independent regions handling the classification of key points in groups (Lemaire et al., 2011). These key points are frequently used in Facial Expressions Recognition (FER) (Happy and Routray, 2015), where mathematical spatial models can define, for example, a distance 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). It provides 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.

Most approaches use spatio-temporal information to classify facial data (Sikdar, 2017) (Yu and Poger, 2017). Different mathematical models for temporal data representation are found in the literature (Erwig et al., 1998) (Lee et al., 2016) (Mahmoud et al., 2014) (Suheryadi and 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 of the models developed and can help to understand the behavior of emotion animation.

Algorithms such as Data Time Warping (DTW), Factor Analysis (FA) and Principal Components Analysis (PCA) can be used to understand the correlation between landmarks and the expressions through their behavior (Mahmoud et al., 2014). Oliveira et al. (Oliveira et al., 2017) use 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 sign definition include the use of landmarks and Principal Component Analysis (PCA) and can be classified by machine learning techniques. This concept will be used to define a facial region during an expression animation as a 4D geometric mesh.

# 3 MODEL FOR AUTOMATIC SYNTHESIS OF FACIAL EXPRESSIONS IN SIGN LANGUAGES THROUGH 3D AVATAR

In the following sections, a novel model for the automatic synthesis of facial expressions in sign languages is presented using the proposed non-manual markers based on regions and Spatio-temporal facial behavior. The NMM model follows the manual parameters format defined for sign language computational models around the world. Using this model, methods for facial behavior analysis was applied to extract 4D meshes that detail the relationship between facial landmarks and base expressions. With this model, an automatic facial expression generation and parser system have been developed that can be integrated into automatic sign language synthesis systems.

### 3.1 Proposed Sign Language Non-manual Markers based on Region Centroid

The model proposed in this section relies on hierarchical modeling, compatible with concepts 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 data-sets that require 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 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.

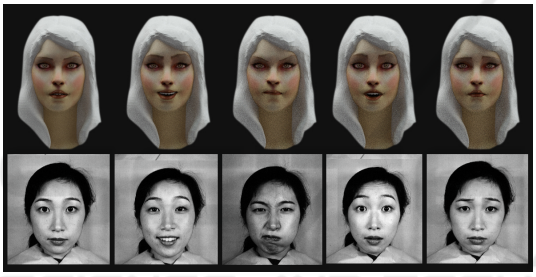

Figure 1: Basic emotions represented in the virtual environment. The top row shows faces built with the 3D mesh, from left to right: Neutral, Joy, Anger, Surprise and Sadness. The bottom row shows the samples of the data-set (Lyons et al., 1998) used as a reference for the base expressions.

A humanoid model was built with 598 facial polygons and a rigging supporting the aforementioned base expressions. The facial model MPEG-4 FP (Kacorri and Huenerfauth, 2014) was used as a points of interest reference in avatar modeling. According to Obaid et al. (Obaid et al., 2010), the main regions for facial expression recognition are: forehead, eyes, cheeks, nose, and mouth.

Other studies using facial regions for the extraction or classification of characteristics argue for 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) (Lv et al., 2015).

Based on these concepts, the facial model for sign language avatar was defined as the merge of the MPEG-4 points of interest controlled by centroids calculated based on the five main facial regions. This setting allows fine control of expression animation and is a new approach in sign language models.

Then, a process for generating the base expression animations was defined using the morph target method (Dailey et al., 2010), where geometric deformations are processed using tracking of points of interest from an input sequence. In the process of tracking and measuring the point of interest, distances were used as a reference two 2D data-sets with images of the base expressions states and representations.

Japanese Female Facial Expression (JAFFE) data-set and the Averaged Karolinska Directed Emotional Faces (AKDEF) data-set provide samples of the base expressions interpreted by more than 20 subjects providing approximately 120 samples (Lyons et al., 1998) (Lundqvist, 1998). These data-sets classify the images using averaged semantic ratings, which define expressions by points of interest values, statistically identifying which emotions each image is and its general intensity(the displacement of the facial muscles and facial points of interest define the expression represents and the value of intensity). Figure 2 shows an example of the AKDEF base where the representation of base expressions uses values for synthesized expression by intensity value.

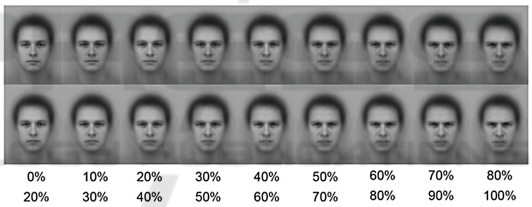


Figure 2: AKDEF data-set sample (Lundqvist, 1998). The images sequences follow the intensity in the representation of the expression (0% for the neutral expression and 100% for the angry base expression), based on the position of the Points of Interest (POI) in the human face. The points of interest is the facial muscles that when displaced indicate the expression or emotion that the face intends to express.

The coordinates for each expression points of interest in face region were defined based on the average values found in the data-sets. Following the semantic values identified in each base expression, controllers were applied to the 3D geometric mesh by deforming the model in order to fit in the average values as shown in Figure 3.

With the morph targets of the applied base expressions, coordinates of specific points on the 3D avatar face, linked to controllers, can be used as predictions of positions for interpolation of expressions. For the input and output parser steps will be used as target descriptive model the CORE-SL, which already implements manual and non-manual sign language parameters and elements, being the most complete model

today.

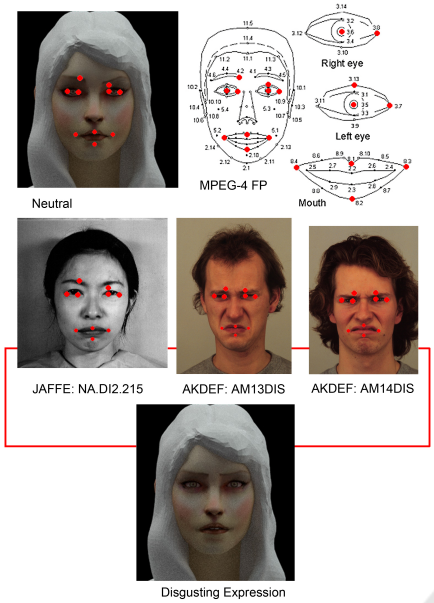


Figure 3: Generation of the synthesized expression based on the subjects of the AKDEF and JAFFE data-sets used. The red dots were defined based on the MPEG-4 and applied to the 3D avatar, and based on the position of the points of interest tracked in the input images, the blend shapes were constructed as animation for the base expressions. In the image, inputs like the three figures displayed (NA.DI2.215, nomenclature used in the JAFFE data-set that defines the subject (NA), expression (DI2) and image number (215), and the AKDEF data-set examples that represent 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.

## 3.2 Neutral Expression Deviation Factor (NEDeF)

The next methods were built aiming to relate each facial region with each base expression. This process is fundamental to identify values for the parameters defined in the previous section and can be used in automatic synthesis in the same way as the manual elements. A metric was 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) as follows:

$$\frac{\sum_{v_{ir} \in R_r} \frac{\left| d_n v_{ir} - d_e v_{ir} \right|}{d_n v_{ir}}}{NV_r} \quad (1)$$

For each expression ($e$) a measure of the distortion relative to the neutral expression ($n$) is computed. This is done, 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 (vertexes placed as MPEG-4 Point of Interest) is defined respective to the centroid. The absolute values of the displacements were used to calculate the NEDeF 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.

Table 1 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 that was extracted in each region.

Table 1: NEDeF of the base expressions for main facial regions, indicating the relation between the facial regions and each expression. The displacements are calculated using centroid position for facial region point of interest.

| Geometrical Comparison of Facial Regions | | | | | |
|---|---|---|---|---|---|
| | Foreh. | Eyes | Cheeks | Mouth | Nose |
| Joy | 0.18 | 1.00 | 0.80 | 0.79 | 0.37 |
| Anger | 0.62 | 0.90 | 0.95 | 0.70 | 0.08 |
| Surp. | 0.36 | 0.46 | 0.00 | 0.97 | 0.00 |
| Fear | 0.25 | 0.48 | 0.67 | 1.00 | 0.09 |
| Sadn. | 1.00 | 0.98 | 0.30 | 0.40 | 0.09 |
| Disg. | 0.73 | 0.05 | 0.66 | 1.00 | 0.00 |

The results presented in Table 1 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) the nose region tends to have a more noticeable change in comparison with the positive emotions (joy and surprise).

The next facial experiment aimed to identify the main components of the face on the emotion synthesis process. This is important to identify the general weight of each region, which will allow prioritizing regions of greater importance in the synthesis of facial animations. For this process the Principal Components Analysis (PCA) and Factor Analysis (FA) algorithms were used, reducing the data dimensions, making it easier to observe the regions with the most relevant influence in the facial expression animations.

The facial landmarks were used as the PCA variables with their Euclidean values ranging from a neutral expression to the extracted synthesized expressions. The covariance matrix was calculated as the eigenvalues and eigenvectors of the average vector of the samples, relating the landmarks and the euclidean displacement. Then the data was rearranged in a Hotelling matrix, used to solve such multivariate testing problems, in order to obtain the facial principal components. Figure4 shows a graph representing the landmarks in two-dimensional space and their representation after applying the PCA algorithm.

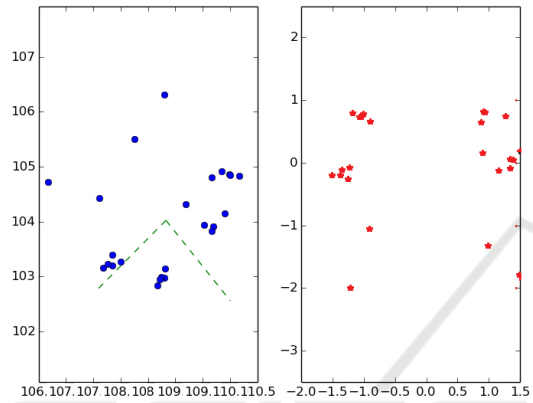


Figure 4: PCA applied to 3D mesh landmarks vector: According to Explained Variance Ratio (EVR), components of Forehead and Mouth regions had the most expressive displacement values. The dotted lines at left show the maximum variance direction of the first and second principal components, and the right image shows the points represented in the new principal component's base.

The Explained Variance Ratio (EVR) was obtained based on the eigenvectors and eigenvalues; it was observed that 58 % of the variance of the data is in the direction of the main 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 Eye region had less expression in the tests, followed by the nose region which had the EVR value lower than 1 %.

In order to support the previous experiments, the Factor Analysis Algorithm (FA) was applied. This algorithm confronts the facial landmarks behavior given by PCA results. The same PCA variables of the previous analysis were used as facial landmarks coordinates in expression animation morph target, observing the weights of their relations.

The objective of applying the FA was to find the co-variance between the regions in the synthesis of facial expression, defining the most relevant regions in the synthesis of each base 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 show that using the five main facial regions of the proposed sign language NMM and the six base expressions of the Ekman's model, it's correct to consider three main factors. The graph in Figure 5 shows the reduced factors based on the eigenvalues, where the absolute values of mouth and nose regions can be reduced in a single factor, as well as values for cheeks and forehead.

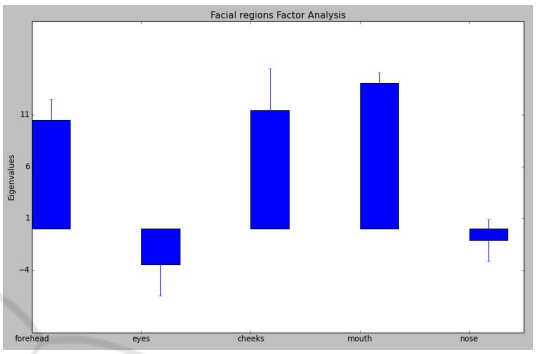


Figure 5: Main factors for facial regions. The five values, from left to right, represents the absolute values for the main facial regions: Forehead, Eyes, Cheeks, Mouth, and Nose.

This means that, in the synthesis of base emotion expressions, using the points of interest landmarks and regions of the proposed 3D avatar facial model, defined in the previous sub-section, we can point out 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 factor with less perceptive displacement. The PCA and FA results reinforce the proposed NEDeF function and point to the more expressive facial regions and centroids in base expression animation. Trajectories of these centroids can define the general facial behavior of the 3D avatar in the sign language synthesis process and can be extracted using Spatio-temporal, or 4D concepts. With the results obtained, these strategies are explored in the next section.

# 4 A MODEL FOR FACIAL LANDMARKS TRAJECTORY BEHAVIOR

The following methods were built aiming at the extraction of the displacement geometric meshes of each controller regarding the centroids of the main facial

regions. These results allow to define the behavior of each controller and generate trajectories that, observed as curves, allow the complete analysis of the synthesis process of emotions and the generation of complex facial expression interpolations.

For this process, the parameters of the NMM for sign language 3D avatar defined in the previous sections were transformed into function curve structures. In order to represent a function curve, the values of the points represented by $t$ are a part of the sequence called Knot Vector and determine the base 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 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 $\alpha$ observed as a time point $t(\alpha)$ as follows:

$$t = p_0, p_1, ...p_n$$

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

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

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

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 considered in the animation generation process, where a lower value of $W$ corresponds to less important regions that can be ignored in the expression synthesis, reducing the computational load.

## 4.1 Spatio-temporal Centroid Extraction

The Spatio-temporal concept used in this work follows the presented by Erwig and Güting (Martin Erwig and Güting, 1998) where the trajectory of a spatial point, based on temporal readings can be observed as a region if its spatial displacement is considered. When the shape of the curve changes, the region displacements is expanded or retracted.

The centroids trajectories along the generated facial expression animation can be represented by dynamic curves, controlling the edges that connect the vertexes landmarks, producing the movement or deformation of facial regions. The relevant coordinates occur in the transition between the neutral expression to the synthesis of one of the six base 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 trajectories of the Facial Expression Landmarks (FEL) of the facial NMM model defined in the previous section were extracted using intervals of 60 frames for each Expression by:

$$Traj = \sum_{i=1}^{n} Cent_{R_1}, [E_0, E_1]$$

When 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 $Cent_{R_1}$, and their displacement. The centroids displacements of the region define their behavior and influence for each expression.

The splines shown in Figure 6 represent the geometric displacement calculated by the Three-dimensional Euclidean Distance (3DED) of the Centroid coordinates $C_{E_1 1}$ observed at 50 $t$s ($t$ relating to an element of the Keyframes vector for the emotion synthesis). Each spline in images defines the Centroid trajectory by the expression, the displacement projection can be considered the 4D data as a trajectory mesh (Le et al., 2011), once represent a geometric controller in a time slice.

Figure 6 shows the Spatio-temporal trajectories of the five main facial regions for the base expressions. Graphically it is possible to observe which region has more spatial variation considering its geometric coordinates and the shape of the curve.

The 3D Euclidean Distance values of each Centroid reading are assigned to the matrix sequenced by the Keyframe of the observed animation, where Euclidean Distance Variance values of a mouth facial region Centroid are displayed and their overall variance value. Based on 3D Euclidean Distances, Table 2 shows the Variance Analysis Matrix for each Centroid in the time slice of the synthesis of each base expression.

Most of the values in 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 variance of 0.00236 and 0.004911 respectively, significant among the matrix average.

The forehead region has a 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 2) indicates that the regions of the eyes with 0.002369 and cheeks with 0.001781 are those that have a greater variance in the displacements also

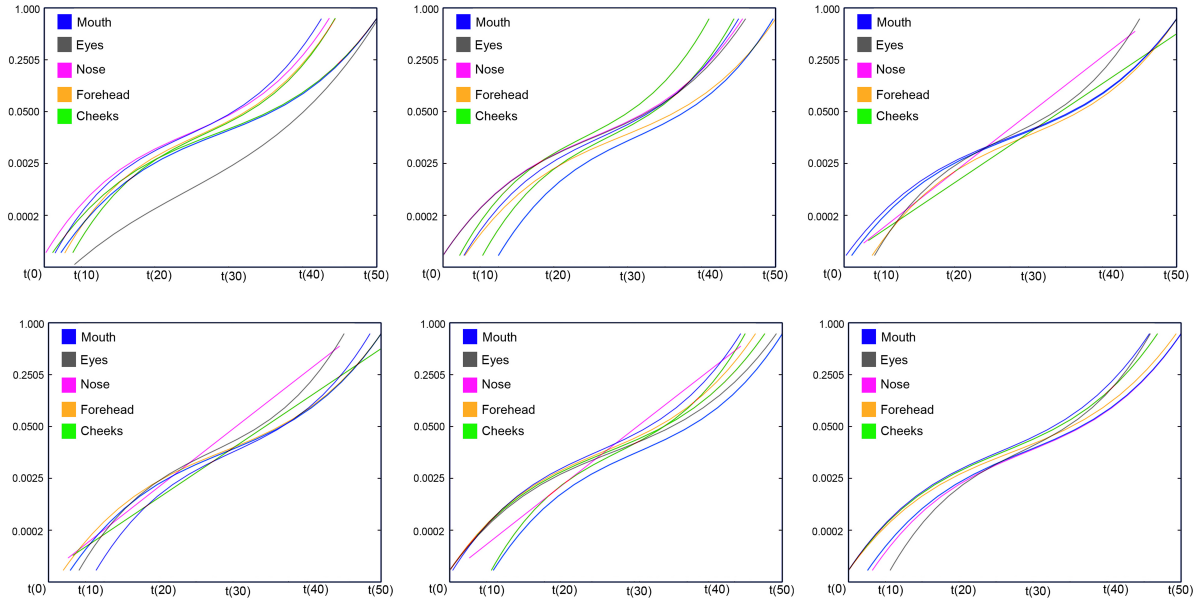


Figure 6: Spatio-temporal Centroid trajectories, from left-top to right bottom: Joy, Anger, Surprise, Sadness, Fear, Disgust expression synthesis. The Spatio-temporal trajectories for each Centroid follow the displacement of 3D Euclidean Distance Variance as shown in Table 2.

Table 2: Centroid trajectory EDVA values. For each base expression, as Joy, Anger, Surprise, Fear, Sadness and Disgusting, the variance value of each facial region is shown (M. for mouth region, C. for cheeks region, N. for nose region, F. for forehead region and E. for eyes region).

| | EDVA values. | | | | | |
|---|---|---|---|---|---|---|
| $C_n$ | Joy | Anger | Surp. | Fear | Sadn. | Disg. |
| M. | .123 | .001 | .002 | .010 | .197 | .182 |
| C. | .163 | .001 | .002 | .002 | .001 | .003 |
| N. | .002 | .000 | .000 | .000 | .000 | .002 |
| F. | .052 | .001 | .076 | .014 | .004 | .011 |
| E. | .006 | .002 | .001 | .001 | .004 | .001 |

reinforcing the values tabulated by the PCA test.

The avatar's mouth presented high values of geometric displacement variance in practically all base expressions, highlighting the disgust expression, with 0.1820 of variance and sadness expression, with 0.1972, 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.

Table 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.

There are many approaches to interpolate curves or Spatio-temporal data where intermediate points are calculated in order to obtain a predicted trajectory that extends the original shape (Gloderer and Hertle, 2010) (T. Jusko, 2016). From a weight parameter, simplifies curves by interpolating duplicated or less relevant edges shortening the number of controller points in the spline (Li et al., 2014), a useful resource for interpolated expressions generation.

The proposed process was applied to create a data-set with different outputs representing interpolated emotions or simplification of base expressions parameters. The Newton Polynomial Interpolation method 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.

In this process, 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 data-sets.

With a list of intermediate values $u$, spatial vertexes displacements are simplified generating a new Spatio-temporal trajectory. With the exported coor-

Table 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 ((M. for mouth region, C. for cheeks region, N. for nose region, F. for forehead region and E. for eyes region).

| | Joy | Anger | Surp. | Fear | Sadn. | Disg. |
|---|---|---|---|---|---|---|
| PCA normalized variance values. | | | | | | |
| M. | .626 | .005 | .011 | .053 | 1.0 | .923 |
| C. | .829 | .009 | .003 | .003 | .005 | .020 |
| N. | .001 | .003 | .002 | .003 | .003 | .002 |
| F. | .267 | .006 | .387 | .071 | .020 | .057 |
| E. | .003 | .012 | .009 | .009 | .024 | .007 |
| 3D EDVA normalized values. | | | | | | |
| M. | .791 | .702 | .970 | 1.00 | .400 | 1.00 |
| C. | .800 | .953 | .000 | .673 | .307 | .667 |
| N. | .376 | .080 | .000 | .095 | .091 | .000 |
| F. | .118 | .628 | .363 | .250 | 1.00 | .730 |
| E. | 1.00 | .900 | .460 | .484 | .980 | .050 |

dinates it is possible to intensify the animations by increasing the variance in the displacements or optimize the generation of the expression 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 morph target by the distance between Centroid and the data matrix using the function that calculates the distance between $C_u$ and $C_j$ in $[C_0, C_{n-1}]$.

The whole process of outputting and calculating base expression animation or expression interpolation has been extracted from a parser process that follows the CORE-SL framework structure. This structure follows an XML notation model where parameters, relationships, and values are defined in order to generate a sign. For the facial model proposed by this work, the parameters are the points of interest and Centroid of each facial region, and their values are the geometric displacement relations defined by the 4D meshes. In summary, the presented generation process is complete and can generate any interpolation variation proposed by the Plutchik Wheel of Emotions, or intensify and alter the base expressions to suit local and regional representations.

# 5 EVALUATION AND DISCUSSION

In order to identify whether the generated expressions are recognizable, two validation tests were applied: an evaluation using K-nearest neighbors algorithm (KNN), Support Vector Machine (SVM) and Machine Learning, and an evaluation collecting feedback from the Deaf community.

## 5.1 Using Machine Learning

Although there are no models that define detailed NMM parameters for sign languages 3D avatar, some models and landmarks are widely used for facial expression recognition (Figure 7), and can be used in order to point out facial behaviors that define emotions and expressions for the virtual environment generation.

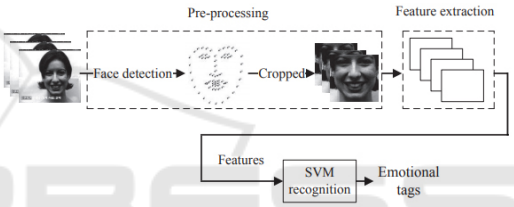


Figure 7: Facial expression recognition using Support Vector Machine (Song et al., 2018).

KNN and SVM are pattern recognition algorithms that use machine learning to classify an input based on a previously trained space of hypothesis $H_d$ to recognize patterns that define classes in a database.

Two facial bases, JAFEE and AKDEF, were used in order to define the facial model in section 3. Both bases were used in SVM training, where characteristics such as eye position, mouth corners, nose, and other spatial characteristics were defined as parameters of variables and the base expressions were defined as labels (Adil et al., ) and (Song et al., 2018).

The KNN algorithm does not learn any hypothesis model being a distance-based strategy of elements by comparing their attributes and defining their dimensional position based on their values.

After training with 2D image bases, 3D Facial Expressions (3DFE) dataset was generated using the model proposed in this work, with 21 examples of synthesized base expressions and some interpolations between them. Although the model generates virtual environment animations as outputs, the final output of each rendered expression can be used as an example and applied to SVM and KNN for labeling, once the coordinates of the landmarks can be identified. In this

way, based on $H_d$, which is adaptable and learns from new classifications, it is possible to label each output of the dataset identifying if the intended expression was generated correctly.

Figure 8 shows the accuracy of Machine Learning algorithms in the labeling of facial expressions generated.

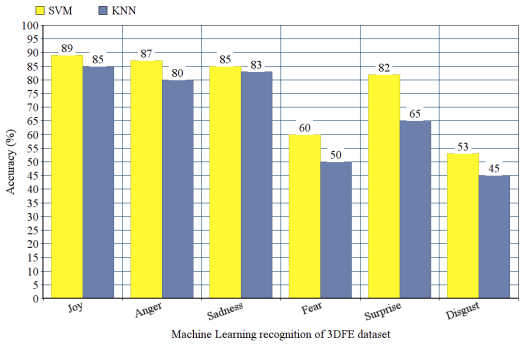


Figure 8: Average accuracy's of 3DFE classification.

It is possible to identify that the KNN algorithm had had less accuracy in classifying the output images, perhaps because its hypothesis space is totally defined by distance calculations. The SVM results were supported by the training stage based on AKDEF and JAFEE databases, and have shown high positive classifications.

The results demonstrate that the expressions generated by the methods proposed in this work are correct and highlight an important first step in the integration of complex expressions in avatars for sign language automatic synthesis systems, and can also be integrated into other sign language systems around the world. A dataset with the base expressions and fifteen interpolations was generated aiming to test the automatic generation process and to present to the deaf community through sign language synthesis systems through 3D avatar.

## 5.2 3DFE Dataset and User Evaluation

A questionnaire containing expressions generated by the synthesis system developed in this work was organized to be presented to a university community of deaf students. The objective was to evaluate if the generated automatic expressions are recognizable by the people, including the interpolations generated through the spatio-temporal geometric meshes. This exercise helps you identify which expressions can have their parameters adjusted for better representation and values for interpolations.

A form containing a collection of 21 synthesized expressions outputs generated using the proposed technique was made available to Sign Language users in order to evaluate the system and its outputs. The form was made available online, and did not require users identification or personal information once it was the system being analyzed and not the people. Users should identify the facial expressions presented, including some interpolations with intensity variations in facial regions, and simplifications by excluding regions with the less relevant *w* parameter.

The application of the form has received 30 anonymous responses. It was reported that 46.7% of the respondents already used software with 3D avatars as sign language interpreters, 10% had already used software that uses 3D avatars but not in the context of sign languages and 43.3% had never used 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. Table 4 presents Spearman's Rank Correlation coefficient and significance applied to the data extracted from the answers feedback. The data below were organized in order to relate the importance of the parameters used to define the outputs in the user's responses.

Table 4: Spearman's rank correlation coefficient applied to validated parameters in the applied formulas. The presented values vary between 1 for a direct relation between the parameters and -1 for inverse relation. 0 represents no relation. The parameters presented are: "B.E." for base expressions, Interpolations (I.), 'N.I.' for unidentified expressions, 'E.' and 'H.' for the number of errors and hits by output image.

| | B.E. | I. | N.I. | E. | H. |
|---|---|---|---|---|---|
| B.E. | 1.00 | -0.75 | -0.33 | -0.09 | 0.91 |
| I. | -0.75 | 1.00 | -0.09 | 0.13 | -0.06 |
| N.I. | -0.33 | -0.09 | 1.00 | 0.15 | -0.51 |
| E. | -0.09 | 0.13 | 0.15 | 1.00 | -0.90 |
| H. | 0.91 | 0.06 | -0.51 | -0.90 | 1.00 |

The positive correlation between the basic expressions and the hits of 0.9 justify 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.

Other variables in Table 4 highlight the strong correlation of interpolated expressions with N.I. responses and the strong correlation between N.I. and Errors. Therefore, the system would be improved if there were image bases with representations of secondary expressions to serve as hypothesis space helping to classify these variations geometrically.

Some relationships can be made between clas-

sifications using artificial intelligence and user responses. Negative expressions had more false positives, especially expressions of fear and disgust. The most successful expressions in both tests were the expressions of joy and anger, as well as sadness. One hypothesis is that these emotions have less cultural variation and are more recognizable regardless of facial and local characteristics. The Deaf community response with the generated expressions was positive as well as the SVM and KNN results; this represents an important step in integrating complex facial expressions into avatars for automatic signal language synthesis, using computational parameters, as it is already done with manual elements.

# 6 CONCLUSION

There is a lacking concerning facial parameter for sign language avatars, despite the importance of expressions in communication. This work presented a proposal of non-manual markers for sign language facial expressions automatic generation following computational descriptive models such as CORE-SL, integrating complex elements for automatic generation of facial expressions unified to manual signals. As main contributions the parameterized computational model that allows the generation of base expressions and secondary interpolations was presented, enabling the representation of any expression based on fine control of facial regions.

The behavior of facial landmarks was also analyzed through Spatio-temporal modeling and application of PCA, FA and EDVA algorithms, which define the correlation between facial regions and the main expressions. These results are important to understand how to computationally generate complex facial expression interpolations.

A dataset with facial expressions generated using the proposed model was created and validated using machine learning algorithms and presented to the deaf community with positive acceptance of the facial expressions and emotions synthesized, The results obtained so far represent a 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. Future works can define relations between base expressions and interpolations and apply the parameters presented in other 3D meshes reinforcing the generality of the model. In addition, it is essential that morphosyntactic elements, defined in each sign language, be integrated into the model.

# ACKNOWLEDGMENTS

This work was financially supported by the São Paulo Research Foundation (FAPESP) (grants #2015/16528-0, #2015/24300-9 and Number 2019/12225-3), and CNPq (grant #306272/2017-2). We thank the University of Campinas (UNICAMP) and Universidade Federal do Paraná (UFPR) for making this research possible.

# REFERENCES

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

Adil, B., Nadjib, K. M., and Yacine, L. A novel approach for facial expression recognition.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hyde, J., Carter, E. J., Kiesler, S., and 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., and Antunes, D. R. (2017). Automatic signwriting generation of libras signs from core-sl. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, IHC 2017, pages 55:1–55:4, New York, NY, USA. ACM.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Neidle, C., Bahan, B., MacLaughlin, D., Lee, R. G., and 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.

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

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

Punchimudiyanse, M. and Meegama, R. (2015). 3d signing avatar for sinhala sign language. In *Industrial and*

*Information Systems (ICIIS), 2015 IEEE 10th International Conference on*, pages 290–295.

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

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

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

Song, N., Yang, H., and Wu, P. (2018). A gesture-to-emotional speech conversion by combining gesture recognition and facial expression recognition.

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

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

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

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

Yu, S. and 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.