Landmark-based Facial Expression Parametrization for Sign Languages Avatar Animation

Diego Addan Gonçalves
Universidade Federal do Paraná
Curitiba, Brazil
diegoaddan@gmail.com

Eduardo Todt
Universidade Federal do Paraná
Curitiba, Brazil
todt@inf.ufpr.br

Débora Pereira Cláudio
Universidade Tecnologica Federal do Paraná
Curitiba, Brazil
deborapclaudio@yahoo.com.br

ABSTRACT

Facial expressions and associates emotions important factors defining the message tone and context information in both spoken and sign language communication. Sign language virtual systems use structural models that define control values specifying body configurations in the animation process, where facial parameters are generally relegated to simple templates or completely neglected. In this work, a facial expression parametrization for avatars is proposed through an procedure that aims to identify the most relevant facial landmarks and emotions in the context of sign languages, in order to enhance automatic sign synthesis systems. An analysis of the influence of landmarks on a geometric mesh based on MPEG-4 model of human face is performed in this research, aiming to identify the principal components and their relationships, in order to allow further optimization of the animation process, supporting faster and lighter avatar animation.

ACM Classification Keywords

I.3.7 COMPUTER GRAPHICS: Three-Dimensional Graphics and Realism

Author Keywords

facial expression, virtual interpreter, avatar animation, sign languages, accessibility, deaf community

INTRODUCTION

Systems that use virtual environments for message transmission are of fundamental importance in modern life. A constant concern in this area is to conceive these systems in such a way to provide persons with special needs easier access to information. [15] [19].

Although sign languages have been documented since the 17th century, their practical definitions and modeling vary locally based on countries legislation [15] [10] [24], most of which being very 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 [22].

There are several Sign Languages synthesis systems developed around the world [12] [3] [10] [20], based on signal synthesis through user interaction. In general, these signal synthesis systems use configuration parameters for hand gestures, besides the body and arms positioning, aiming fidelity between the virtual and the real representations [12] [21]. It is possible to observe the predominant lack or weak representation of facial expressions, although these features provide message context and intensity modifiers [6]. The accurate transmission of a message really needs facial expression as a feeling modifier or as context supplement for the raw gesture information. In [17] 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. Besides enhancing communication, facial expressions are considered very important in order to a robot or avatar being accepted by humans [24].

The objective of this work it is to search for a landmark-based parametrization of facial expressions for use in a signal language synthesis system, in order to enhance the animation process of current systems, supporting faster and lighter avatar animation. The main facial points relevant to the expressions representation on a 3D mesh are identified together with their spatial trajectory when performing movements associated to emotions.

The facial expression parametrization is proposed through the identification of the most relevant facial landmarks and emotions in the context of sign languages. An analysis of the influence of each landmark on a 3D geometric mesh based on MPEG-4 model of human face is performed, aiming to identify the principal components and their relationships. Thus, the first step in this process is the identification of the principal components in the deformation of facial regions while building expressions. As a result, the most affected facial areas in the avatar during the synthesis process are characterized. This makes possible to generate the interpolation of synthesized emotions in independent areas, so that the less important components for the expression synthesis may be discarded, reducing the total computational cost.

In Figure 1 an overview of proposed system is shown. The Facial Parameters block contains a representation of the fa-

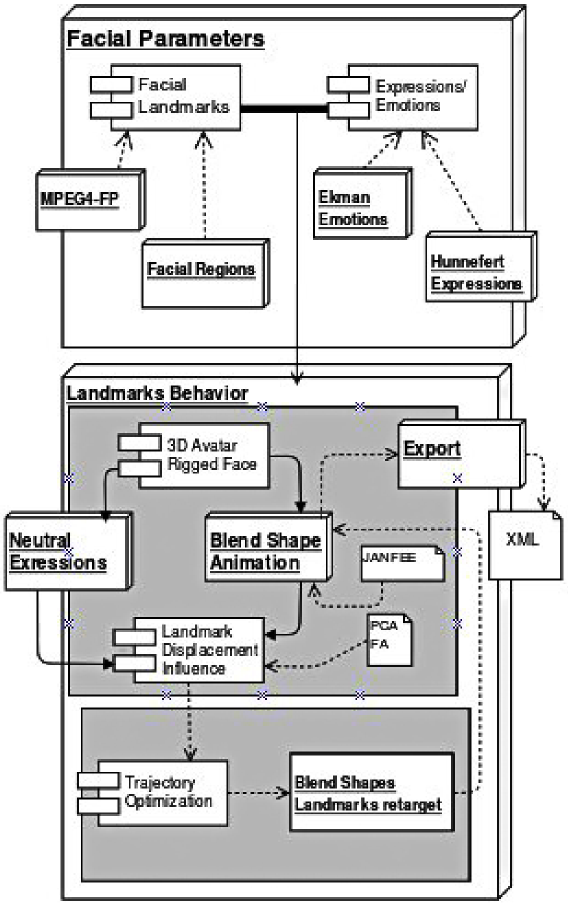

Figure 1. Overview of the animation process. There are two blocks: the Facial Parameters block and the Landmarks Behaviour. The first one has the facial landmarks associated to regions of interest to be animated as well as a representation for emotions and expressions to be reproduced. The Landmarks behaviour is responsible for the animation generation.

cial landmarks on the 3D mesh, obtained from the MPEG-4 model and the manual definition of the regions of interest, as well a representation for emotions and expressions to be reproduced. The Landmarks Behaviour block generates the animation, exporting the commands as an XML file, based on the neutral expressions over the rigged face. The animation produces movements corresponding to the landmarks selected by a principal component analysis previously performed on the facial animations modelled. These concepts will be covered in section 2. From this step, it is be possible to optimize the process of interpolation of expressions, through the analysis of the behaviour of the landmarks during the movements associated to expressions.

FACIAL LANDMARKS FOR SIGN LANGUAGES

This section is divided into two main topics. The first reviews works concerning the use of 3D avatars oriented to the representation of sign languages in virtual environments, while the the second focuses on the features used to represent facial expressions and the computational representation of emotions.

Virtual Interpreters for Sign Languages Systems

Current approaches for obtaining more realistic animations of virtual interpreters, called avatars, use motion capture with tracking of landmarks in videos of real interpreters or extract depth coordinates using 3D sensors [3] [10]. The main challenge to obtain a plastic animation relies in the details, in the sense that the corresponding synthesis process demands greater computational cost than that required by a coarse animation, considering that much more specific parameters will be rendered.

The avatar motions may be generated by controllers associated with the 3D mesh using techniques such as Blend Shapes or Morph Targets [10] [5]. Some models use notation of sign languages, as Signwriting and HamNoSys, extending them for representing characteristics of non-manual signal elements, using terms such as symmetry, hand position, rotation, location, among other information [13]. These models make parallel between signs and their descriptive notations, where symbols are used to represent the terms and movements used by the speaker [15]. Also, such models typically doesn't cover facial expressions.

In this direction, systems that use virtual interpreters need a strong focus on parametrized representation of facial expressions [8]. Thus, the main issues in the process of animation synthesis using avatars is a well defined set of parameters in addition to a fine detail control on the geometric mesh.

Facial Expression Features

According to [17], 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. The face contains less noticeable but important variations. Those parameters can be defined as intensities of expressions interpolation, displacement of landmarks and message context [25].

Modulation in the mouth, eyebrows and other regions of interest in the face can change the meaning of an expression in sign languages, beside details such as variation in volume and timing in spoken languages [7]. 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 [6] [7].

The main 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 [1] [14]. Ekman's model [23] identifies base expressions as *Anger*, *Fear*, *Sadness*, *Surprise*, *Desgusting* and *Joy*. Other representations of emotions in general are identified as a combination of the previously mentioned.

MPEG-4 Feature Points, or Facial Definition Parameters, is a broadly used standard characterizing points of interest in the face [4] [10] [2]. It has 84 points mapped on a model face with a neutral expression, including areas such as tongue, lips, teeth, nose and eyes, with points distributed along the perimeter of these regions, mainly at the corners.

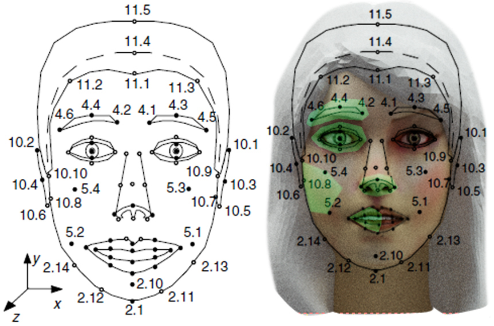

Figure 2. Facial Regions, in green, and facial action parameters based in MPEG-4 model. The defined regions include key areas in face on synthesis process, that are: Forehead, Mouth, Cheek, Eyes and Nose and consider geometric symmetry aiming to optimize the experiments.

This model can be used as a basis for setting facial controllers associated with a 3D geometric mesh, and also supports the use of additional features, such as variations in texture [9] or variation in color representation of the face [1].

FACIAL REGIONS AND COMPONENTS

This section presents an analysis based on the distortions of polygons bounding the relevant facial regions affected by the movements performed during the manifestation of an expression.

Facial Parameters Vector

A humanoid model was built with 598 facial polygons with a rig, supporting the aforementioned basic expressions, subsection 2.2. The facial model MPEG-4 FP [11] was used as a reference in the avatar modeling, also the most representative regions of the face were set as forehead, eyes, cheeks, nose and mouth, as depicted in Figure 2. Theses regions were defined on the basis of experiments shown in [18], describing regions for facial deformation modelling.

The deformations analysed in this work were built using Blend Shapes based on the Japanese Female Facial Expression (JAFFE) database, which provides examples of the basic emotions interpreted by 10 Japanese subjects [16]. This database classifies the images using a semantic rating, which defines values for expressions, statistically identifying to which emotions each image corresponds and its intensity, in general terms.

The landmarks were tracked in the neutral expression images using the defined facial parameters. Blend shapes were animated based on the values relating the displacement of the corresponding landmarks with the points of the geometric mesh.

In the following, an experiment aiming to identify the most affected regions and principal components on 3D mesh in the process of synthesis of facial expressions is shown. The region deformations are evaluated based on the analysis of the spatial displacement of the landmarks locations, where the landmarks are reference points associated with the bounding polygons of the regions of interest.

Influence of Expressions in Facial Regions

The metric proposed to measure the influence of the expressions on the mesh is based on the region deformation. This is defined as the normalized averaged relative spatial displacement of the landmarks, given by the following equation:

$$\frac{\sum_{v_{ir}\in R_r}\frac{|d_n v_{ir} - d_e v_{ir}|}{d_n v_{ir}}}{NV_r}$$

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 landmark associated with this region. A second normalization is computed considering the total number of landmarks defined for the region NV_r. The distances were taken as absolute values because the distortion of the regions is assumed to be additive.

Table 1 shows the distances 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.

The results of this first step (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 as well 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). It is important to mention that, based on Plutchick's model, emotions can be interpolated [23], thus, intermediary expressions considered in the model can be obtained. Also the characteristics can be modulated by an intensity factor.

The next step was the identification of the the principal components of the face on emotion synthesis. For this, a Principal Component Analysis was performed, with five samples taken during the expressions' movements, related to the facial regions. The Euclidean distances of the landmarks' positions to the centroid of the respective region defined the values to be compared to the neutral expression. The covariance matrix was calculated as the eigenvalues and eigenvectors of the average vector of the samples.

Landmarks Geometrical Comparison					
	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

Table 1. Displacement influence on the 3D mesh of facial regions on synthesis of an expression.

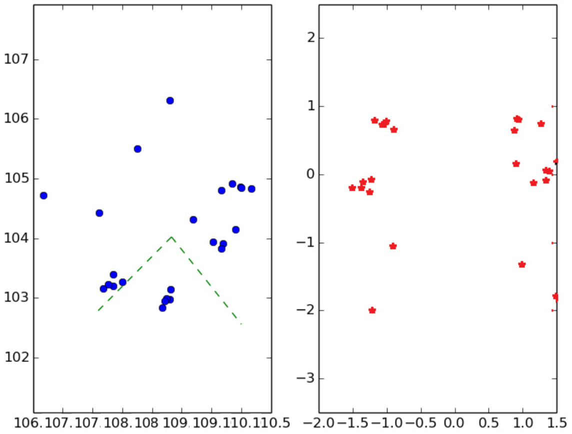

Figure 3. PCA applied to 3D mesh landmarks vector

Thereafter the data was rearranged in a Hotelling Transposed Matrix, in order to obtain their Principal Components. Figure 3 depicts a graph representing the components in two dimensions and its representation after applying the PCA algorithm. The dotted lines at left show the maximum variance directions 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, 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 Eye region had less expression in the tests, followed by the nose region with had the EVR value lowest than 1 %.

These results enable an landmark behavior discussion, giving values for each trajectory in geometric space. In future work, an optimized animation process can be develop based on these extracted values.

In order to support the previous experiments, Factor Analysis Algorithm was applied. This experiment makes easy to understand the facial landmarks behavior. For this, same PCA variables were used as facial regions and facial expressions, observing the weights of their relations.

The Figure 4 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 extracted from the centroids displacements from the neutral position to the Euclidean coordinates after the expression synthesis. The centroids are calculated from the landmarks 3D coordinates of each facial region.

The objective of the FA application was to find the covariance between the regions in the execution of an expression using the weight of the relation between the data of the geometric mesh.

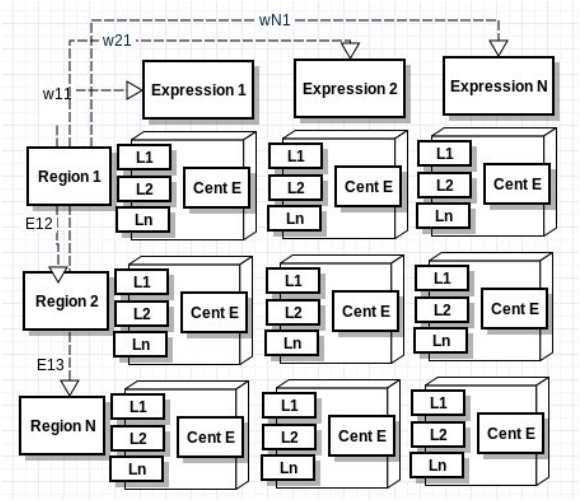

Figure 4. Data model for FA Algorithm, where centroid coordinate absolute values are calculated between facial region and expressions based on their weights.

The simplification of extracted factors should be useful in identifying which regions are most expressive in the observed synthesis process.

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 graphic in Figure 5 shows the reduced factors based on the eigenvalues, where it is possible to observe by the absolute values that regions 1 and 3 can be reduced in a single factor, as well as values 2 and 5.

This means that, in the synthesis of base emotion expressions, using the landmarks based on the MPEG4-FP model and the regions defined in section 3, we can point that the Forehead and Cheeks regions had an 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 calculated as a single factor with less perceptive displacement.

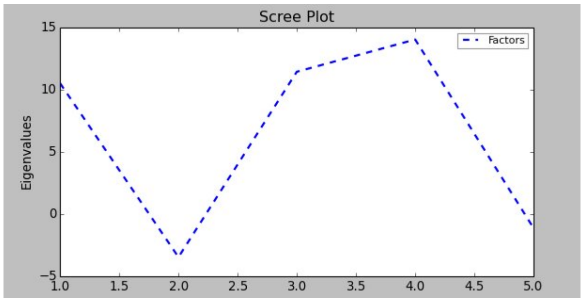

Figure 5. Main factors for facial regions. The five values, from left to right, represents the absolute values for regions: Forehead, Eyes, Cheeks, Mouth and Nose

Observing the results of the PCA and FA tests we can conclude that the landmarks with greater expressiveness in the synthesis of emotions can have their reduced dimensions based on the proximity of displacement in Euclidean space. The next objective to define more precisely the facial landmarks behavior is to analyze the interpolation process of emotions, as well as their coordinates in a temporal model.

CONCLUSIONS

This paper presents a preliminary study intended to identify the relationship between facial regions deformations and the emotion synthesis in order to, in future work, optimize the representation of facial expressions for sign languages, reducing computational cost of the animation process. Based on the performed tests it is possible to propose techniques to optimize computational cost of interpolating blend shapes at points where the distortions are more frequent and spatially related. Also, in future work it is intended to evaluate the relationship between points using algorithms like Dynamic Time Warping, where the temporal relationships during expression sequences are considered.

ACKNOWLEDGEMENT

The authors would like to acknowledge and thank Brazilian National Counsel of Technological and Scientific Development (CNPq) for financial support in this work.

REFERENCES

1. Mohammed H. Alkawaz and Ahmad H. Basori. 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)*. ACM, New York, NY, USA, 143–152. DOI: http://dx.doi.org/10.1145/2407516.2407555

2. Koray Balci. 2004. Xface: MPEG-4 Based Open Source Toolkit for 3D Facial Animation. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '04)*. ACM, New York, NY, USA, 399–402. DOI: http://dx.doi.org/10.1145/989863.989935

3. Jose Bento, Ana P. Claudio, and Paulo Urbano. 2014. Avatars on Portuguese sign language. In *Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on*. 1–7. DOI: http://dx.doi.org/10.1109/CISTI.2014.6876959

4. Yozra Bouzid, Oussama El Ghoul, and Mouhamed Jemni. 2013. Synthesizing facial expressions for signing avatars using MPEG4 feature points. In *Information and Communication Technology and Accessibility (ICTA), 2013 Fourth International Conference on*. 1–6. DOI: http://dx.doi.org/10.1109/ICTA.2013.6815304

5. Matthew N. Dailey, Carrie Joyce, Michael J. Lyons, Miyuki Kamachi, Hanae Ishi, Jiro Gyoba, and Garrison W Cottrell. 2010. *Evidence and a computational explanation of cultural differences in facial expression recognition*. Emotion, Vol 10(6). "874–893" pages. DOI: http://dx.doi.org/10.1037/a0020019

6. Ain S. Elons, Menna Ahmed, and Hwaidaa Shedid. 2014. Facial expressions recognition for arabic sign language translation. In *Computer Engineering Systems (ICCES), 2014 9th International Conference on*. 330–335. DOI: http://dx.doi.org/10.1109/ICCES.2014.7030980

7. Matt Huenerfauth, Pengfei Lu, and Andrew Rosenberg. 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)*. ACM, New York, NY, USA, 99–106. DOI: http://dx.doi.org/10.1145/2049536.2049556

8. Jennifer Hyde, Elizabeth J. Carter, Sara Kiesler, and Jessica K. Hodgins. 2016. Evaluating Animated Characters: Facial Motion Magnitude Influences Personality Perceptions. *ACM Trans. Appl. Percept.* 13, 2, Article 8 (Feb. 2016), 17 pages. DOI: http://dx.doi.org/10.1145/2851499

9. Asim Jan and Hongying Meng. 2015. Automatic 3D facial expression recognition using geometric and textured feature fusion. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 05. 1–6. DOI: http://dx.doi.org/10.1109/FG.2015.7284860

10. Hernisa Kacorri. 2015. TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation. (2015). Retrieved September 9, 2017 from http://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1402&context=gc_cs_tr.

11. Hernisa Kacorri and Matt Huenerfauth. 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)*. ACM, New York, NY, USA, 261–262. DOI: http://dx.doi.org/10.1145/2661334.2661387

12. Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, and Mackenzie Willard. 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 &#38; Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 147–154. DOI: http://dx.doi.org/10.1145/2700648.2809860

13. Sandeep Kaur and Maninder Singh. 2015. Indian Sign Language animation generation system. In *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*. 909–914. DOI: http://dx.doi.org/10.1109/NGCT.2015.7375251

14. Maja Kocon. 2014. Facial expressions modeling for interactive virtual environments. In *Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On*. 744–747. DOI: http://dx.doi.org/10.1109/MMAR.2014.6957448

15. Vincenzo Lombardo, Cristina Battaglino, Rossana Damiano, and Fabrizio Nunnari. 2011. An Avatar-based Interface for the Italian Sign Language. In *Complex, Intelligent and Software Intensive Systems (CISIS), 2011 International Conference on*. 589–594. DOI: http://dx.doi.org/10.1109/CISIS.2011.97

16. Michael J. Lyons, Shigeru Akemastu, Miyuki Kamachi, and Jiro Gyoba. 1998. *Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition*. 200–205 pages.

17. Carol Neidle, Benjamin Bahan, Dawn MacLaughlin, Robert G. Lee, and Judy Kegl. 1998. Realizations of syntactic agreement in American sign language: Similarities between the clause and the noun phrase. *Studia Linguistica* 52, 3 (1998), 191–226. DOI: http://dx.doi.org/10.1111/1467-9582.00034

18. Mohammad Obaid, Ramakrishnam Mukundan, Mark Billinghurst, and Catherine Pelachaud. 2010. Expressive MPEG-4 Facial Animation Using Quadratic Deformation Models. In *Computer Graphics, Imaging and Visualization (CGIV), 2010 Seventh International Conference on*. 9–14. DOI: http://dx.doi.org/10.1109/CGIV.2010.11

19. Malinda Punchimudiyanse and Gayan N. Meegama. 2015. 3D signing avatar for Sinhala Sign language. In *Industrial and Information Systems (ICIIS), 2015 IEEE 10th International Conference on*. 290–295. DOI: http://dx.doi.org/10.1109/ICIINFS.2015.7399026

20. Rabindra Ratan and Béatrice S. Hasler. 2014. Playing Well with Virtual Classmates: Relating Avatar Design to Group Satisfaction. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing (CSCW '14)*. ACM, New York, NY, USA, 564–573. DOI: http://dx.doi.org/10.1145/2531602.2531732

21. Thomas Rieger. 2003. Avatar Gestures. In *WSCG 2003, International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, Vol. 2. 379–386.

22. Cassia G. Sofiato and Lucia H. Reily. 2014. Brazilian sign language dictionaries: comparative iconographical and lexical study.. In *EducaÃğÃčo e Pesquisa, 40(1) 109-126. https://dx.doi.org/10.1590/S1517-97022014000100008*.

23. Wioleta Szwoch. 2015. Model of emotions for game players. In *Human System Interactions (HSI), 2015 8th International Conference on*. 285–290. DOI: http://dx.doi.org/10.1109/HSI.2015.7170681

24. Ewerton J. Wantroba and Roseli A. F. Romero. 2015. An Interactive Question-Answer System with Dialogue for a Receptionist Avatar. In *2015 12th Latin American Robotics Symposium and 2015 3rd Brazilian Symposium on Robotics (LARS-SBR)*. 360–365. DOI: http://dx.doi.org/10.1109/LARS-SBR.2015.54

25. Karl Wiegand. 2014. Intelligent Assistive Communication and the Web As a Social Medium. In *Proceedings of the 11th Web for All Conference (W4A '14)*. ACM, New York, NY, USA, Article 27, 2 pages. DOI: http://dx.doi.org/10.1145/2596695.2596725