TrueDepth measurements of facial expressions: Sensitivity to the angle between camera and face

Esselink, L.; Oomen, M.; Roelofsen, F.

DOI
10.1109/ICASSPW59220.2023.10193107

Publication date
2023

Document Version
Final published version

Published in
IEEE ICASSPW 2023 workshop proceedings

2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) | 979-8-3503-0261-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSPW59220.2023.10193107

TRUEDEPTH MEASUREMENTS OF FACIAL EXPRESSIONS: SENSITIVITY TO THE ANGLE BETWEEN CAMERA AND FACE

Lyke Esselink, Marloes Oomen, Floris Roelofsen

University of Amsterdam,
Amsterdam, the Netherlands
{l.d.esselink; m.oomen2; f.roelofsen}@uva.nl

ABSTRACT

Facial expressions play an important role in communication, especially in sign languages. Linguistic analysis of the exact contribution of facial expressions, as well as the creation of realistic conversational avatars, especially sign language avatars, requires accurate measurements of the facial expressions of humans while engaged in linguistic interaction. Several recent projects have employed a TrueDepth camera to make such measurements. The present paper investigates how reliable this technique is. In particular, we consider the extent to which the obtained measurements are affected by the angle between the camera and the face. Overall, we find that there are generally significant, and often rather substantial differences between measurements from different angles. However, when the measured facial features are highly activated, measurements from different angles are generally strongly correlated.

1. INTRODUCTION

Facial expressions play an important role in communication. This is especially clear in sign languages, where facial expressions can contribute to lexical content, convey grammatical information (e.g. whether a sentence is a statement or a question), and relay affective content (e.g. whether the speaker is satisfied or not) [1, 2]. In spoken languages, grammatical information and affective content can also be conveyed by facial expressions, in tandem with prosody [3, 4, 5].

Analysis of the exact linguistic contribution of facial expressions in signed and spoken languages, as well as the creation of realistic conversational avatars, especially sign language avatars, requires accurate measurements of the facial expressions of humans while engaged in linguistic interaction.

Several recent projects have employed TrueDepth cameras to make such measurements [6, 7, 8, 9, 10, 11]. TrueDepth cameras are built into recent iPhone and iPad models, primarily for identification purposes. They are able to automatically recognize the face of the device's owner, giving them access to the device without the need to enter a passcode. Evidently, identification requires high fidelity. This means that the measurements made by TrueDepth cameras are exceedingly precise and discriminative. In principle, this makes them suitable to obtain fine-grained measurements of facial expressions for the purpose of linguistic analysis and avatar synthesis. Another attractive aspect of TrueDepth cameras is that they are relatively affordable and highly portable compared to other depth-sensing equipment.

We gratefully acknowledge financial support from the Netherlands Organization for Scientific Research (NWO, grant number VI.C.201.014).

However, this new technique to measure facial expressions for the purpose of linguistic analysis and avatar synthesis also raises important methodological questions. How reliable are the measurements? How reproducible are they? The present paper takes a first step in addressing these questions. Specifically, we investigate to what extent the measurements of a TrueDepth camera are affected by the horizontal and vertical angle between the camera and the measured face.

The paper is organized as follows. Section 2 provides more elaborate background and motivation, Section 3 describes our method, and Section 4 presents the results. Finally, Section 5 draws general conclusions, highlights several limitations of the present study, and suggests avenues for future work.

2. BACKGROUND AND MOTIVATION

2.1. Traditional methods based on video

Most research so far on the role of facial expressions in signed and spoken languages is based on video data. Such data, however, is two-dimensional and therefore never fully captures the actual physical reality that it represents, which is three-dimensional. Furthermore, important details are sometimes not visible on video footage because of a limited frame rate, limited resolution, motion blur, or occlusion (e.g. a hand in front of the face). Ideally, researchers would be able to base their analysis on data that captures facial expressions in a format that stays closer to the original, with less inherent transformation (3D to 2D), compression (frame rate, resolution), and noise (blur, occlusion).

To enable linguistic analysis, video data is usually first annotated. The annotation of facial expressions, making use of the Facial Action Coding System (FACS) [12, 13] or similar coding schemes [14, 15], is a notoriously laborious process. Moreover, even when done with great care, manual annotation has some inescapable limitations. It is inherently *subjective* (two annotators may disagree as to whether an eyebrow is raised or neutral), *not robustly reproducible* (a single annotator may label an eyebrow as raised one day, and the same eyebrow as neutral six months later), and inherently *categorical* (an eyebrow can be labeled as raised or neutral, perhaps 'half raised', but not 'raised to degree 0.35') while in reality eyebrow raise and other facial features are quantitative/continuous variables, not categorical ones—so in the annotation phase the data is further 'compressed', losing part of the original information. Ideally, researchers would be able to obtain detailed representations of facial expressions in a way that is less laborious, not subjective, reproducible, and quantitative rather than categorical (meaningful categories may be identified in a later stage of analysis, but should not be imposed on the researchers from the start).

2.2. Recent approaches using keypoint detection

Recent work by Kimmelman et al. [16, 17, 18] partly addresses the limitations of manual annotation of facial features, building on [19, 20, 21]. Kimmelman et al. use OpenFace face recognition software [22] to automatically detect a signer's eyebrows and eye-corners, and compute a degree of eyebrow raise/lowering in terms of the distance between these. This method to extract degrees of eyebrow raise/lowering from video data is automatic, objective, and quantitative. However, there are still some limitations. First, measurements of relevant facial features like brow raise are *indirect* and *not robustly reproducible*. OpenFace detects facial keypoints. Features have to be derived from distances between keypoints, but this cannot be straightforwardly done in a reliable way because these distances depend on the distance and angle between the camera and the signer's face (as discussed in [17]), which are impossible to keep constant across and even within recordings. Second, the proposed method still takes 2D *video data* as its starting point. So, while this body of work makes an important first step in addressing the limitations of manual annotations, it does not address the issues of inherent transformation, compression and noise associated with video data.

2.3. Recent approaches using a TrueDepth camera

Several recent projects [6, 7, 8, 9, 10, 11] aim to overcome these issues by using a depth sensing camera instead of an ordinary video camera to measure facial expressions. Specifically, they make use of a TrueDepth camera, which is built into recent models of the iPhone and the iPad, in combination with the Live Link Face application by Epic Games. A TrueDepth camera projects 30,000 infrared dots on the face and measures the distances between these dots. Based on these measurements, a detailed 3D representation of the face is computed. From this 3D representation, 52 facial blendshapes are derived, as well as 9 rotational features (3 for the head and 3 for each eye). We focus here on the blendshapes, which include, for instance, BROWDOWN, BROWOUTERUP, EYEWIDE, EYESQUINT, CHEEKSQUINT, and MOUTHFROWN (in each case there is in fact a separate blendshape for the left and the right side of the face). Blendshape coefficients are measured at 60 frames per second. Each blendshape coefficient is a value between 0 to 1, indicating the degree of engagement of the facial feature at hand.

Unlike OpenFace, which detects keypoints based on video input, this method bypasses the main issues associated with video data, and moreover directly measures facial features that are of interest for linguistic research as opposed to keypoint coordinates, which first have to be translated into feature coefficients, something which, as mentioned above, cannot always be done in a straightforward way.

Some of the recent projects cited above use TrueDepth measurements primarily for the purpose of linguistic analysis [11]; others use the measurements to drive speaking avatars [6, 7, 8] or signing avatars [9, 10]. While these developments show much promise, both for linguistic analysis and for avatar synthesis, it is important to inquire into the potential limitations of this new technique to measure facial expressions. As a first step, we focus here on how sensitive the measurements of a TrueDepth camera are to the angle between the camera and the face.

3. METHOD

3.1. Data collection

Three participants (one male, two female) were instructed to display a sequence of facial expressions. We simultaneously measured these facial expressions with five TrueDepth cameras (C0,...,C4). All five cameras were placed at a horizontal distance of 63cm from the participant's face. C0, which we refer to as the 'reference camera', was placed straight in front of the participant's face. C1 was placed 23cm above C0, C2 33cm below C0, C3 40cm to the right of C0 (from the participant's perspective), and C4 40cm to the left of C0. All data was collected under the same lighting conditions.

Fig. 1: Experimental setup

The sequence of expressions that participants displayed consisted of brow raises (3x), brow lowerings (3x), a scrunched up face with intense cheek and eye squint (3x), eye blinks (3x), mouth shrugs (3x), mouth frowns (3x), pressed lips (3x), and funneled lips (3x). Finally, participants pronounced the sentence "The quick brown fox jumps over the lazy dog".

All recordings were made using the free Live Link Face app made by Epic Games for the iPhone. Recordings were synchronised using NTP-based timecodes (an option that is available in the app).

3.2. Data pre-processing

Data pre-processing was carried out in Python. Frames recorded by different cameras were first matched according to their timecodes. To ease interpretability of the results, all blendshape measurements were multiplied by 100. The dataset was restricted to frames with timecodes for which all cameras contributed a frame without NULL blendshape values. We removed frames with 'diverging values' (defined below) for one of the head rotation features, HeadPitch, HeadRoll, and HeadYaw, because rotation of the head affects the angle between the camera and the face, which is meant to be kept constant during each recording. A value for a rotation feature x measured by camera y in recording z was considered 'diverging' if it was more than 2.5 standard deviations away from the mean of all values for x measured by y in recording z. If a frame was removed for one camera, corresponding frames from other cameras were also removed. Finally, we excluded from the analysis blendshapes related to eye-gaze direction (EYEIN, EYEOUT, EYEUP, EYEDOWN for both eyes), the jaw (JAWLEFT, JAWRIGHT), the tongue (TONGUEOUT), and some related to the mouth (MOUTHLEFT, MOUTHRIGHT), because these were not significantly engaged in the facial expressions that participants displayed. The final dataset for analysis comprised 114.851 frames, each involving measurements for 39 blendshapes.

3.3. Data analysis

The analysis consisted of pairwise comparisons between measurements of the reference camera, C0, with the other cameras (i.e., C0 vs C1, C0 vs C2, C0 vs C3, C0 vs C4).

Table 1: Effects of vertical angle going up, C0 vs C1

(a) High Activation

Blendshape	Intercept	Effect	%Effect	Corr	#Frames
EYESQUINTLEFT	43.7	16.9 *	39	0.76	2814
EYESQUINTRIGHT	43.8	16.9 *	39	0.75	2818
EYEWIDELEFT	51.1	5.9 *	12	0.92	2196
EYEWIDERIGHT	51.0	5.9 *	12	0.92	2204
MOUTHFROWNLEFT	51.4	−3.8 *	7	0.94	4050
MOUTHFROWNRIGHT	51.3	−6.2 *	12	0.94	3820
MOUTHSHRUGLOWER	58.1	−1.5 *	3	0.84	4048
MOUTHSHRUGUPPER	55.4	−6.2 *	11	0.88	3454
BROWDOWNLEFT	58.6	−0.6	1	0.94	2748
BROWDOWNRIGHT	58.6	−0.7	1	0.94	2746
BROWINNERUP	64.9	−8.9 *	14	0.91	2442
BROWOUTERUPLEFT	56.6	−5.0 *	9	0.93	2362
BROWOUTERUPRIGHT	56.6	−5.0 *	9	0.93	2366
CHEEKSQUINTLEFT	33.0	4.0 *	12	0.90	2922
CHEEKSQUINTRIGHT	36.5	5.8 *	16	0.91	2890

(b) Low Activation

Intercept	Effect	%Effect	Corr	#Frames
7.7	19.3 *	251	0.32	27578
7.7	19.3 *	251	0.32	27574
6.4	2.8 *	44	0.61	11416
6.4	2.9 *	45	0.60	11436
6.2	4.6 *	74	0.35	12166
7.8	0.2	3	0.33	11732
10.9	14.3 *	131	0.58	26344
7.3	9.9 *	136	0.67	26932
7.8	2.6 *	33	0.73	21200
7.7	2.5 *	32	0.73	21056
4.6	3.6 *	78	0.68	27196
9.2	−7.3 *	79	0.30	6590
9.2	−7.3 *	79	0.29	6710
5.0	7.2 *	144	0.64	27470
5.1	8.4 *	165	0.67	27502

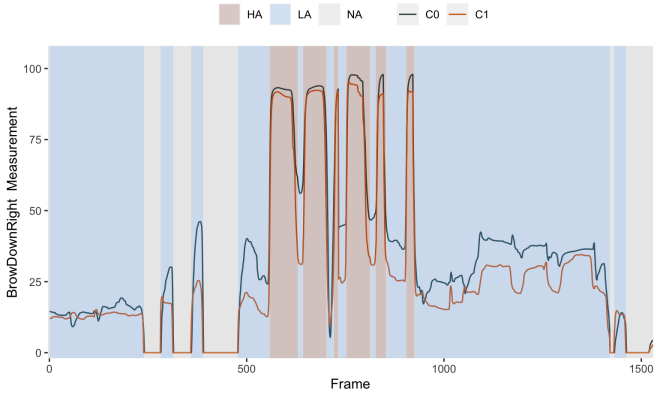

Fig. 2: Activation level classification

For each blendshape b, each pair of cameras c_i and c_j, and each recording r, we made a distinction between *High Activation* (HA), *Low Activation* (LA), and *No Activation* (NA) frames. b was classified as *activated* at frame f in recording r according to camera c_i if the value of b as measured by c_i exceeded a minimal threshold $\theta_{c_i}^{\text{min}}$, which we set to 3. Similarly, b was classified as *highly activated* at f according to c_i if the value of b as measured by c_i exceeded the threshold $\theta_{c_i}^{\text{high}}$, defined as the mean of all values of b measured by c_i in r plus 0.5 times the standard deviation of these values.

When comparing measurements of a blendshape b by two cameras c_1 and c_2, we classified a frame f as HA if b was highly activated at f according to both cameras, as LA if it was not HA but still activated or highly activated according to at least one camera, and as NA otherwise.

Only HA and LA frames were further analyzed, NA frames were disregarded. Fig. 2 shows an example classification of the first 1500 frames in one recording for the comparison of measurements by C0 and C1 of the blendshape BROWDOWNRIGHT.

For each pair of cameras, each blendshape, and each activation level (HA/LA), we built a linear mixed effects model using the lmer function from the lme4 package in R [23]. For each model, we specified BLENDSHAPEVALUE as the independent variable, CAMERA as a fixed effect (with the reference camera C0 coded as 0 and the other camera coded as 1), and PARTICIPANT and RECORDING as random effects. In cases where this resulted in a singular fit, we computed a simplified model without RECORDING as random effect. Finally, we calculated Pearson's correlation coefficients for blendshape values measured by the two cameras.

The analyzed data set and all analysis scripts are publicly available as supplementary materials via Github [24].

4. RESULTS

4.1. Effects of vertical angle going up: C0 versus C1

We first consider the effects of vertical angle 'going up'. That is, we compare the measurements of our reference camera, C0, with those of the upper camera, C1. The results are given in Table 1. For reasons of space, we restrict ourselves here to 15 blendshapes, which have been argued to be particularly relevant for linguistic analysis and synthesis of facial expressions [11]. Results for the other blendshapes are given in the supplementary materials and show the same overall pattern.

We first consider HA frames. The *Intercept* column reports, for each blendshape, the Intercept of the fitted linear model, which corresponds to the mean of all measured values for that blendshape by the reference camera C0. In the column *Effect*, we report the main effect of CAMERA, which amounts to the mean difference between the measurements for that blendshape by C0 and C1, respectively. Stars (*) indicate that this difference is significant for all blendshapes except BROWDOWNLEFT and BROWDOWNRIGHT.

Besides knowing whether the mean difference in measurement between the two cameras is significant, it is also of interest to know how large this mean difference is relative to the mean of all C0 measurements for that blendshape. We express this as a percentage, |Effect / Intercept $*100$|, in the column *%Effect*. We see that the percentage differences are low for some blendshapes but quite high for others, ranging between 1 and 39 (mean = 13.1; std = 11.4).

The column *Corr* provides Pearson's correlation coefficients for all blendshapes. These are generally very high, ranging between 0.75 and 0.94 (mean = 0.90; std = 0.06).

Table 2: Effects of horizontal angle going right, C0 vs C3

(a) High Activation

Blendshape	Intercept	Effect	%Effect	Corr	#Frames
EYESQUINTLEFT	45.7	15.0 *	33	0.77	5270
EYESQUINTRIGHT	46.8	9.1 *	19	0.87	5070
EYEWIDELEFT	56.6	−33.5 *	59	0.67	2518
EYEWIDERIGHT	57.9	−23.7 *	41	0.77	2678
MOUTHFROWNLEFT	54.8	−15.9 *	29	0.92	6060
MOUTHFROWNRIGHT	50.7	−11.3 *	22	0.94	7642
MOUTHSHRUGLOWER	59.7	−10.6 *	18	0.68	7796
MOUTHSHRUGUPPER	56.2	−14.8 *	26	0.63	6692
BROWDOWNLEFT	55.1	3.7 *	7	0.87	6254
BROWDOWNRIGHT	54.4	3.3 *	6	0.87	6344
BROWINNERUP	64.9	−10.9 *	17	0.91	3336
BROWOUTERUPLEFT	67.8	−14.1 *	21	0.80	2688
BROWOUTERUPRIGHT	66.8	−12.7 *	19	0.80	2754
CHEEKSQUINTLEFT	37.9	−14.3 *	38	0.81	4884
CHEEKSQUINTRIGHT	39.0	−9.0 *	23	0.88	4934

(b) Low Activation

Intercept	Effect	%Effect	Corr	#Frames
8.3	24.6 *	296	0.42	51036
8.2	18.4 *	224	0.45	51236
10.9	−10.2 *	94	0.35	17500
10.6	−9.7 *	92	0.37	17422
7.9	−2.2 *	28	0.30	35162
8.3	−1.5 *	18	0.41	31564
11.4	1.9 *	17	0.58	48510
7.6	0.7 *	9	0.52	49518
8.5	7.8 *	92	0.69	43688
8.5	7.1 *	84	0.66	43068
6.2	−2.4 *	39	0.59	35444
10.0	−9.6 *	96	0.47	13176
9.7	−9.6 *	99	0.46	13290
5.2	1.1 *	21	0.53	51278
5.2	2.6 *	50	0.56	51372

Finally, in the column *#Frames* we report the number of frames that were taken into account. This number ranges from 2196 to 4048, meaning that the analysis for each blendshape was based on a reasonable number of frames.

We now turn to the results for LA frames, given in Table 1b. There are a couple of striking differences between the results for LA frames and those for HA frames. The percentage differences between the two cameras are much higher for LA frames, ranging from 3 to 251 (mean = 103.0; std = 75.9). The correlation coefficients, on the other hand, are much lower for LA frames, ranging from 0.29 to 0.73 (mean = 0.52; std = 0.18).

4.2. Effects of vertical angle going down: C0 versus C2

Next, we consider the effects of vertical angle 'going down', comparing C0 with C2. Overall, the effects are similar to the effects of vertical angle 'going up'. For reasons of space, we defer detailed tables with results per blendshape for the current Section and Section 4.4 to the supplementary materials. For HA frames, the percentage difference between the two cameras ranges from 0 to 37 (mean = 16.7; std = 12.4). The correlation coefficients range between 0.80 and 0.95 (mean = 0.90; std = 0.05). For LA frames, percentage differences are again much higher, ranging from 11 to 198 (mean = 64.9; std = 49.9); and correlation coefficients much lower, ranging between 0.06 and 0.75 (mean = 0.45; std = 0.21).

4.3. Effects of horizontal angle going right: C0 versus C3

To determine the effects of horizontal angle 'going right' we compare C0 to C3. For HA frames, the percentage differences range from 6 to 59 (mean = 25.2; std = 13.5), and the correlation coefficients range from 0.63 to 0.94 (mean = 0.81; std = 0.10). For LA frames, the percentage differences range from 9 to 296 (mean = 83.9; std = 80.1), and the correlation coefficients range from 0.30 to 0.69 (mean = 0.49; std = 0.11); Table 2 provides detailed statistics per blendshape.

4.4. Effects of horizontal angle going left: C0 versus 4

Finally, to determine the effects of horizontal angle 'going left', we compare C0 to C4. For HA frames, the percentage differences range from 7 to 41 (mean = 19.2; std = 10.4) and the correlation coefficients range between 0.74 and 0.93 (mean = 0.82; std = 0.06). For LA frames, the percentage differences range from 5 to 199 (mean = 58.9; std = 58.5) and the correlation coefficients range between 0.05 and 0.49 (mean = 0.36; std = 0.14).

5. DISCUSSION AND CONCLUSION

Two general patterns emerge from our results. First, for HA frames, while displacement of the camera in any direction (up, down, left, right) generally has a significant and often substantial effect on measured blendshape values (with mean percentage differences between 13 and 25 percent), the different measurements are generally highly correlated (mean correlation coefficients between 0.81 to 0.90).

Second, measurements for LA frames are generally much less reliable than those for HA frames, exhibiting much higher percentage differences and lower correlation coefficients between cameras.

These findings are relevant for any work making use of TrueDepth cameras for linguistic analysis or avatar synthesis. This work needs to take into account that the angle between camera and face can substantially affect the measured blendshape values, although for HA frames measurements from different angles are strongly correlated.

The present study is only a first step in a broader inquiry into the prospects and pitfalls of TrueDepth measurements of facial expressions for linguistic analysis and avatar synthesis. It has several methodological limitations which may be overcome in future work. For instance, it is unknown whether the patterns we found generalize to a larger and more diverse set of participants. Moreover, while participants were of different heights, they all sat on the same stool while being recorded. The camera-tripods were not adjusted to different heights. Future studies may avoid this potential confound.

Besides methodological limitations, the present study evidently has a limited scope as well. One particularly important question that needs to be addressed in future work is to what extent the *distance* between the camera and the face, as opposed to the *angle*, affects the measured blendshape values.

6. REFERENCES

[1] Roland Pfau and Josep Quer, "Nonmanuals: their grammatical and prosodic roles," in *Sign Languages*, Diane Brentari, Ed., pp. 381–402. Cambridge University Press, 2010.

[2] Annika Herrmann and Nina-Kristin Pendzich, *Nonmanual gestures in sign languages*, pp. 2149–2162, De Gruyter Mouton, Berlin, 2014.

[3] Ravindra J Srinivasan and Dominic W Massaro, "Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English," *Language and Speech*, vol. 46, no. 1, pp. 1–22, 2003.

[4] Joan Borràs-Comes and Pilar Prieto, "Seeing tunes: The role of visual gestures in tune interpretation," *Laboratory Phonology*, vol. 2, no. 2, pp. 355–380, 2011.

[5] Marisa Cruz, Marc Swerts, and Sónia Frota, "The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties," *Laboratory Phonology*, vol. 8, pp. 24, 2017.

[6] Ryosuke Miyawaki, Monica Perusquia-Hernandez, Naoya Isoyama, Hideaki Uchiyama, and Kiyoshi Kiyokawa, "A data collection protocol, tool and analysis for the mapping of speech volume to avatar facial animation," in *International Conference on Artificial Reality and Telexistence Eurographics Symposium on Virtual Environments (2022)*. 2022, The Eurographics Association.

[7] Jonathan Ehret, Andrea Bönsch, Lukas Aspöck, Christine T Röhr, Stefan Baumann, Martine Grice, Janina Fels, and Torsten W Kuhlen, "Do prosody and embodiment influence the perceived naturalness of conversational agents' speech?," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 4, pp. 1–15, 2021.

[8] Liyang Chen, Zhiyong Wu, Jun Ling, Runnan Li, Xu Tan, and Sheng Zhao, "Transformer-S2A: Robust and efficient speech-to-animation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7247–7251.

[9] Daniël Vink, "A modern approach to realistic facial animations in signing avatars – a qualitative assessment of the MetaHuman avatar," Master thesis, University of Amsterdam, 2022.

[10] Le Luo, Dongdong Weng, Guo Songrui, Jie Hao, and Ziqi Tu, "Avatar interpreter: Improving classroom experiences for deaf and hard-of-hearing people based on augmented reality," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–5.

[11] Lyke Esselink, "Computer vision and machine learning for the analysis of non-manual markers in biased polar questions in Sign Language of the Netherlands," Master thesis, University of Amsterdam, 2023.

[12] Paul Ekman, W.V. Friesen, and J.C. Hager, "Facial action coding system (FACS)," *A Human Face, Salt Lake City*, 2002.

[13] Nina-Kristin Pendzich, *Lexical nonmanuals in German Sign Language: Empirical studies and theoretical implications*, De Gruyter, 2020.

[14] Marloes Oomen, Tobias de Ronde, and Floris Roelofsen, "A procedure for annotating non-manual markers in question sentences in sign languages," Poster presented at North East Linguistics Society (NELS 53), 2023.

[15] Naomi Nota, James P Trujillo, and Judith Holler, "Facial signals and social actions in multimodal face-to-face interaction," *Brain Sciences*, vol. 11, no. 8, pp. 1017, 2021.

[16] Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova, "Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study," *PloS one*, vol. 15, no. 6, 2020.

[17] Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman, "Using computer vision to analyze non-manual marking of questions in KRSL," in *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, 2021, pp. 49–59.

[18] Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman, "Functional data analysis of non-manual marking of questions in Kazakh-Russian Sign Language," in *Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, 2022.

[19] Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle, "Recognition of nonmanual markers in American Sign Language (ASL) using non-parametric adaptive 2D-3D face tracking," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 2414–2420.

[20] Bo Liu, Jingjing Liu, Xiang Yu, Dimitris Metaxas, and Carol Neidle, "3D face tracking and multi-scale, spatio-temporal analysis of linguistically significant facial expressions and head positions in ASL," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 4512–4518.

[21] Anna Puupponen, Tuija Wainio, Birgitta Burger, and Tommi Jantunen, "Head movements in Finnish Sign Language on the basis of motion capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls," *Sign Language and Linguistics*, vol. 18, no. 1, pp. 41–89, 2015.

[22] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[23] Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker, *lme4: Linear mixed-effects models using Eigen and S4*, 2014, R package version 1.1.

[24] Lyke Esselink, Marloes Oomen, and Floris Roelofsen, *TrueDepth measurements of facial expressions: Sensitivity to the angle between camera and face - Supplementary materials*, 2023, https://github.com/LykeEsselink/LiveLinkFace_Evaluation.