Measuring the Perception of Facial Expressions in American Sign Language Animations with Eye Tracking

Hernisa Kacorri[1], Allen Harper[1], and Matt Huenerfauth[2]

[1]The City University of New York (CUNY)
Doctoral Program in Computer Science, The Graduate Center,
365 Fifth Ave, New York, NY 10016 USA
hkacorri@gc.cuny.edu, aharper@gc.cuny.edu

[2]The City University of New York (CUNY)
Computer Science Department, CUNY Queens College
Computer Science and Linguistics Programs, CUNY Graduate Center
65-30 Kissena Blvd, Flushing, NY 11367 USA
matt@cs.qc.cuny.edu

Abstract. Our lab has conducted experimental evaluations of ASL animations, which can increase accessibility of information for signers with lower literacy in written languages. Participants watch animations and answer carefully engineered questions about the information content. Because of the labor-intensive nature of our current evaluation approach, we seek techniques for measuring user's reactions to animations via eye-tracking technology. In this paper, we analyze the relationship between various metrics of eye movement behavior of native ASL signers as they watch various types of stimuli: videos of human signers, high-quality animations of ASL, and lower-quality animations of ASL. We found significant relationships between the quality of the stimulus and the proportional fixation time on the upper and lower portions of the signers face, the transitions between these portions of the face and the rest of the signer's body, and the total length of the eye fixation path. Our work provides guidance to researchers who wish to evaluate the quality of sign language animations: to enable more efficient evaluation of animation quality to support the development of technologies to synthesize high-quality ASL animations for deaf users.

Keywords: American Sign Language, accessibility technology for people who are deaf, eye tracking, animation, evaluation, user study.

1 Introduction

Over 500,000 people in the U.S. use American Sign Language (ASL), a separate language from English, with a distinct word order, linguistic structure, and vocabulary [17]. For various educational reasons, deaf and hard-of-hearing students perform, on average, lower than their hearing peers on tests of English reading comprehension [20-21]; these students therefore have difficulty with text on curriculum materials,

captioning, or other media. While it is possible to use videos of actual human signers in educational content or websites, animated avatars are more advantageous for several reasons in these contexts. If the information is frequently updated, it may be prohibitively expensive to re-film a human performing ASL, thus leading to out-of-date information. Computer synthesized animations allow for frequent updating, automatic generation or machine translation, animation flexibility, and collaboration of multiple authors to script a message in ASL. Thus, virtual human characters have been favored by sign language synthesis researchers and many educational-system developers. For example, Adamo-Villani et al. [1-2] investigated digital lessons annotated with ASL animation and signing avatars to improve the mathematical abilities of deaf pupils, Vcom3D [3] focused on sign language software tools for early education curriculum, and Karpouzis et al. [15] proposed an educational platform for learning sign language.

Relatively few sign language animation synthesis systems have been developed, due to challenging linguistic aspects of ASL. Signers use facial expressions and head movements to communicate essential information during ASL sentences, and state-of-the-art sign language animations systems do not yet handle facial expressions sufficiently to produce clear and understandable animations. Our lab has recently focused on modeling and synthesizing facial expressions. To evaluate our models, we typically ask native ASL signers to view our animations and then answer comprehension and subjective Likert-scale questions [8–11][16]. The challenge is that signers may not consciously notice a facial expression during an ASL passage [10][14], and some facial expressions affect the meaning of ASL sentences in subtle ways [14], thereby making it difficult to invent stimuli and questions that effectively probe a participant's understanding of the information conveyed specifically by the signer's face.

In this paper, we analyze native ASL signers' perception of ASL animations with and without facial expressions, and videos of a human signer. In a prior study [13], we experimentally evaluated ASL animations with and without facial expressions, using videos of a human signer as an upper baseline. The participants answered subjective and comprehension questions and their eye movements were recorded via an eye-tracker to investigate whether their eye movements can reveal the quality of the animations being evaluated. We found that when viewing videos, signers spend more time looking at the face and less frequently move their gaze between the face and body of the signer, compared to when viewing animations. We also found that the fixation time on the face and the frequency of gaze transitions between the face and the hands was significantly correlated with the subjective scores participants assigned to the animations. Thus, there is potential for eye-tracking to serve as a complementary or alternative method of evaluating ASL animations.

A limitation of this prior study was that we did not observe any significant correlation between these two metrics and participants reporting having noticed a particular facial expression nor their comprehension questions scores. In this paper, we present a second, deeper analysis of the data with more fine-grained Areas of Interest (AOIs) such as the upper face and the lower face of the human or animated signer in the stimuli. This new study also considers a new metric, called Total Trail Distance, which is the aggregated distance between fixations normalized by the stimuli duration.

2 Eye-Tracking and Related Work

The eye tracking literature has been previously surveyed by several authors [6][12][19]. The main benefit of eye tracking for human-computer interaction studies is that it delivers a detailed record of position and timing as subjects gaze at visual stimuli; and it does so in both an unmediated and unobtrusive manner that precludes the use of interruptive methods such as Talk-Aloud protocols [5]. In prior work, eye tracking has been used to record the eye movements of deaf participants who viewed live or video-recorded sign language performances; we are not aware of any prior studies using eye tracking to evaluate sign language *animations*.

For instance, Cavender et al. [4] explored the feasibility of presenting sign language videos on mobile phones. This study evaluated the understandability of sign language when displayed at different sizes and video-compression rates. Participants were eye-tracked while viewing the videos and then answered evaluation questions. They found that participants' gaze transitioned away from the signer's face during fingerspelling, hand movement near the bottom of the screen, or when the signer pointed to locations outside the video. The participants' total trail distance was shorter for the video stimuli that received the highest subjective scores; and the mouth region of the signer received the highest fixation counts.

Muir and Richardson [18] performed an eye tracking study to explore how native British Sign Language (BSL) signers employ their central (high-resolution) vision and peripheral vision when viewing BSL videos. Their earlier studies had suggested that signers tend to use their central vision on the face of a signer, and they tend to use peripheral vision for hand movements, fingerspelling, and body movements. In [18], native BSL signers watched three videos that varied in how visually challenging they were to view: (1) close-up above-the-waist camera view of the signer with no fingerspelling or body movement, (2) distant above-the-knees view of the signer with use of some fingerspelling, (3) distant above-the-knees view of the signer with use of fingerspelling and body movements. Proportional fixation time was calculated over the following five AOI's: upper face, lower face, hands, fingers, upper body, and lower body. Results indicated that detailed signs and fingerspelling did not accumulate large proportional fixation time, indicating that participants used their peripheral vision to observe these aspects of sign language video. In all three videos, the AOIs on the face region received the most proportional fixation time: 88%, 82%, 60% respectively. In contrast, Video 3 included upper body movement, and participants spent more time looking at the upper body of the signer. Comparing sub-regions of the face, during video 1, participants looked at the upper face 72% and lower face 16%, but during video 2 (more distant view of the signer), they looked at the upper face 47% and lower face 35%. Both these results are of interest to our current study because they indicate that participant's gaze will likely shift under conditions of sign language videos that have lower clarity (i.e., signer is more distant from the camera), in an effort to search for the AOI with the most useful and visible information. This indicates that studying proportional fixation time on the face might be a useful way to analyze eye-tracking data when participants are viewing sign language videos (or animations) of different quality.

Emmorey et al. [7] conducted an eye tracking experiment to explore the differences in eye movement patterns between native and novice ASL signers. It was hypothesized that novice signers would have a smaller visual field from which to extract information from a signer. In turn, this would lead to: less time fixating on the singer's face, more fixations on the lower mouth and upper body, and more transitions away from the face to the hands and lower body. Unlike the previous studies, [7] used live signing performances that presented two stories constructed with differing amounts of fingerspelling and use of locative classifier constructions (signs that convey spatial information, investigated in our prior work [9]). The goal of the study was to induce more transitions in novice signers due to their restricted perceptual span. The results showed that both novice and native signers displayed similar proportional fixation times (89%) on the face. In contrast to this pattern, novice signers spent significantly more time fixating on the signer's mouth than native signers, who spent more time fixating on the signer's eyes. It was also observed that neither novices nor native signers made transitions to the hands during fingerspelling, but did make transitions towards classifier constructions.

3 Prior Work, Eye Tracking Metrics, and Hypotheses

In prior work [13], we conducted a user-study in which native ASL signers watched animations of ASL (of varying levels of quality) while an eye-tracker recorded them. In that prior study, we examined whether there was a relationship between the quality of the stimuli and participants' proportional fixation time on the face of the signer or the number of "transitions" between the face and the hands of the signer. We examined the following hypotheses [13]:

- H1: There is a significant difference in native signers' eye-movement behavior between when they view videos of ASL and when they view animations of ASL.
- H2: There is a significant difference in native signers' eye-movement behavior when they view animations of ASL with some facial expressions and when they view animations of ASL without any facial expressions.
- H3: There is a significant correlation between a native signer's eye movement behavior and the scalar subjective scores (grammatical, understandable, natural) that the signer assigns to an animation or video.
- H4: There is a significant correlation between a native signer's eye movement behavior and the signer reporting having noticed a facial expression in a video or animation.
- H5: There is a significant correlation between a native signer's eye movement behavior and the signer correctly answering comprehension questions about a video or animation.

We found that, when viewing videos, signers spend more time looking at the face and less frequently move their gaze between the face and body of the signer in support of H1. We also found that H3 was supported for animations, there were significant correlations between these two eye-tracking metrics and participants' responses to subjective evaluations of animation-quality. However, the results for H2, H4, and

H5 were inconclusive and H3 was only partially supported for videos. A limitation of our earlier study was that we did not distinguish between the upper (above nose) and lower face of the signer in the video. Muir and Richardson [18] had distinguished between these parts of the face, and they found changes in proportional fixation time on the face of signers when the visual difficulty of videos varied. Since many grammatically significant ASL facial expressions consist of essential movements of the eyebrows, in this paper, we separately analyze the upper and lower face.

Since Cavendar et al. [4] had found a relationship between the path length of eye gaze the quality of videos of human signers, in this paper, we also measure the "trail length" of the fixations of the participants eye gaze when watching stimuli. Given that Emmorey et al. [7] found that less skilled signers transitioned their gaze to the hands of the signer more frequently, we predict that there will be longer "trail lengths" of the eye gaze in our lower-quality animations, which are harder to understand.

4 User Study

In [13], participants viewed short stories in ASL of three versions: a high-quality "video" of a native ASL signer, a medium-quality animation with facial expressions based on a "model," and a low-quality animation with no facial expressions. The stories were scripted and performed by native signers, and each story was produced in all three versions. The video size, resolution, and frame-rate for all stimuli were identical. Participants responded to three types of questions after viewing a story: First, they answered Likert-scale subject questions about the grammatical correctness, ease of understanding, and naturalness of movement. Next, they answered on a Likert-scale as to whether they noticed a facial expression during the story. Finally, they answered four comprehension questions about the content of the story. The comprehension questions were designed so that wrong answers would indicate that the participants had misunderstood the facial expression displayed [14].

In [13], only two areas of interest (AOIs) were considered for the analysis of participants' eye gazing behavior: "Face" and "Hands". In this paper, we divided the "Face" AOI to "Upper Face" and "Lower Face" AOI based on the signers' nose-tip height. Fig. 1 illustrates these areas of interest for the animations of the virtual character (with or without facial expressions) and for the videos of the human signer. Note that during a small fraction of time signers may move their hands in front or close to their face thus the two AOIs could overlap. Currently, this is handled by a simplifying assumption that the face should take precedence, and that is why the "Hands" AOI has an irregular shape to accommodate the "Face" AOI. We believe that this limitation in our analysis had a minimal effect on the results obtained, given that the signer's hands do not overlap with the face during the vast majority of signing.

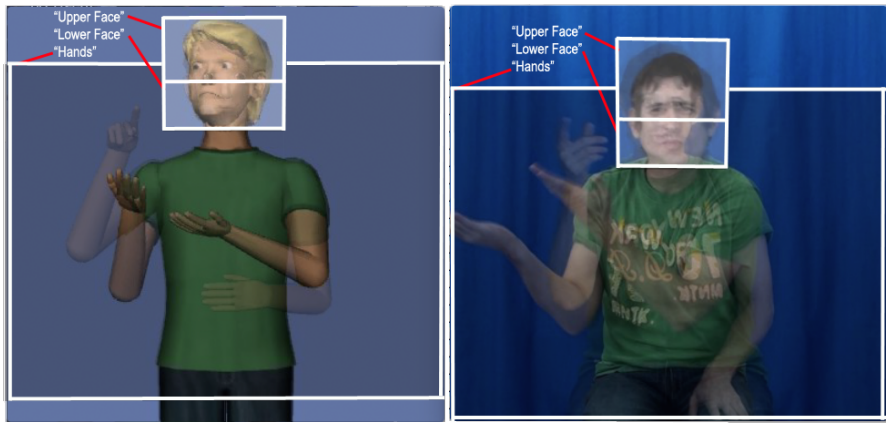


Fig. 1: Screen regions for the upper face, lower face, and hands AOIs.

The AOIs were defined identically for all animations (with and without facial expressions). While the area (width x height) of the face AOIs were preserved, the vertical-horizontal ratio was slightly different for human videos: The human would often bend forward slightly, therefore the region of the screen where his head tend to occupy is a little lower compared to the animated character. So, we set the nose-tip line slightly lower for the human signer; to preserve fairness, we kept the area of the "Upper-Face" and "Lower-Face" AOIs as similar as possible between the animated character and human signer (97.6% for the upper and 102.6% for the lower portion).

As described in detail in [13], eleven ASL signers were recruited for the study and were recorded by an eye tracker as they watched the animations and videos. Eye tracking data was excluded from analysis if the eye tracker equipment determined that either of the following conditions had occurred for over 50% of the time of the video or animation: (a) the eye-tracker could not identify the participant's head and pupil location or (b) the participant looked away from the computer screen.

5 Results

This section presents the results of the eye-tracking data analysis from the eleven participants, and the discussion is structured around three types of metrics:

- Transition frequency (i.e., the number of transitions between pairs of AOIs, divided by story duration in seconds) between the upper-face AOI and the hands-body AOI and between lower-face AOI and hands-body AOI.
- Proportional Fixation Time on the upper-face AOI or on the lower-face AOI (i.e., the total time of all fixations on the AOI, divided by story duration)
- Time-Normalized Total Trail Length (i.e., the sum of the distances between all of the participant's fixations, divided by the story duration in seconds).

Transition frequencies are displayed as a box plot in Fig. 2, with the min/max values indicated by whiskers, quartiles by the box edges, and median values by a center line (not visible in Fig. 2(a) because the median value was zero). On the basis of

Kruskal-Wallis tests, significant differences are marked with stars ($p<0.05$). The three groups displayed include "Video" of a human signer, a "Model" animation with facial expressions, and a "Non" animation with no facial expressions. There was a significant difference between the transition frequency between upper-face and body-hands, comparing Video and Non animations.

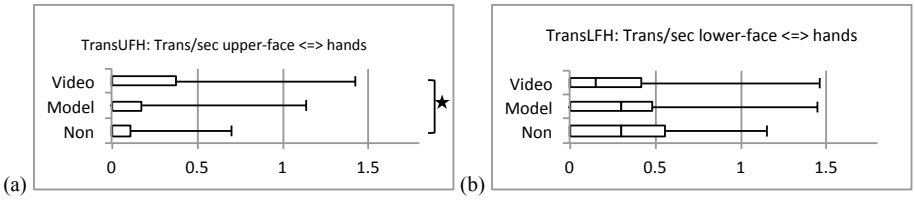


Fig. 2. Transitions per second between: (a) the hands-body AOI and the upper-face AOI ("TransUFH") and (b) the hands-body AOI and the lower-face AOI ("TransLFH").

In order to better understand where participants were looking during the videos or animations, we also calculated the proportion of time their eye fixations were within the upper-face or lower-face AOIs; the results are shown in Fig. 3. In this case, a significant difference was shown between Video and both types of animation (Model and Non) when considering the lower-face AOI in Fig. 3(b). Only the pair Video vs. Non was significantly different when considering the upper-face AOI in Fig. 3(a).

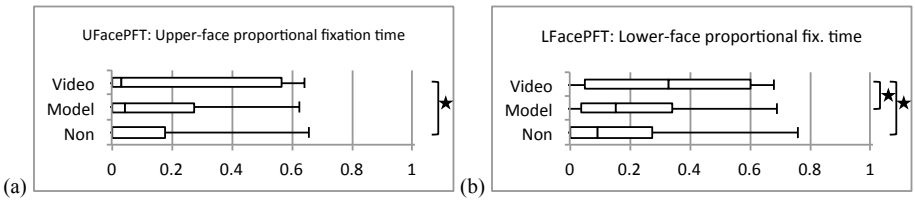


Fig. 3. Proportional fixation time on: (a) the upper-face AOI (labeled as "UFacePFT") and (b) the lower-face AOI (labeled as "LFacePFT").

Since H2 was not supported, Model and Non were grouped together when calculating correlations to investigate Hypotheses H3, H4, and H5. Spearman's Rho was calculated, with significant correlations ($p<0.05$) marked with stars in Fig. 4. Overall, the metrics using the upper-face AOI were more correlated to participants' responses to questions about the animations; most notably, Fig. 4(a) shows significant correlations between the proportional fixation time on the upper-face AOI ("UFacePFT") and participants' responses to Likert-scale subjective questions in which they were asked to rate the grammaticality, understandability, and naturalness of movement of the animations. This result supports hypothesis H3 for animations, but not for Videos of human signers. No significant correlations were found between the eye metrics and the other types of participants' responses: questions about whether they noticed facial expressions and comprehension questions about the information content of videos or animations. Based on these results, H4 and H5 were not supported.

(a)

Spearman's Rho (* if p < 0.05)	UFacePFT Video	UFacePFT Anim.	TransUFH Video	TransUFH Anim.
Grammatical	0.149	* -0.340	0.166	* -0.305
Understandable	0.056	* -0.346	0.161	-0.145
Natural Movement	0.073	* -0.402	0.191	* -0.213
Notice Face Expr.	0.060	-0.101	0.058	-0.099
Comprehension	-0.001	-0.086	-0.064	-0.090

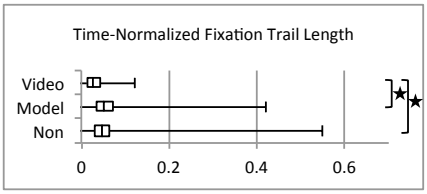
(b)

Spearman's Rho (* if p < 0.05)	LFacePFT Video	LFacePFT Anim.	TransLFH Video	TransLFH Anim.
Grammatical	0.087	-0.092	0.189	-0.090
Understandable	0.147	-0.156	0.217	-0.660
Natural Movement	0.093	* -0.215	* 0.277	-0.029
Notice Face Expr.	0.023	-0.239	0.198	-0.003
Comprehension	-0.018	-0.047	-0.030	0.027

Fig. 4. Correlations between participants responses (rows) and eye metrics (columns), including proportional fixation time and transition frequency for upper-face and lower-face.

The final eye metric considered in this paper is the time-normalized total trail length, which is shown in Fig. 5. There was a significant difference between Video and both types of animation (Model and Non) in Fig. 5(a), further supporting hypothesis H1. The correlations between this metric and the participants' responses are shown in Fig. 5(b). This metric had significant correlations with the greatest number of types of participant responses, as indicated by the stars in Fig. 5(b). While there was still no support for hypotheses H4 or H5, based on the results in Fig. 5(b), hypothesis H3 was supported for both videos of human signers and animations of virtual humans.

(a)



(b)

Spearman's Rho (* if p < 0.05)	TrailLen Video	TrailLen Anim.
Grammatical	* 0.369	0.096
Understandable	* 0.378	* 0.227
Natural Movement	* 0.440	* 0.292
Notice Face Expr.	0.119	0.084
Comprehension	-0.035	-0.035

Fig. 5. Fixation trail length for each type of stimulus (a) and correlations to responses (b).

6 Discussion and Future Work

This paper has identified how eye-tracking metrics are related to participants' judgments about the quality of ASL animations and videos. We have investigated and characterized differences in participants' eye-movement behavior when watching human videos or virtual-human animations of ASL. The results of our user study are useful for future researchers who wish to measure the quality of ASL videos or animations: eye-tracking metrics that can serve as complimentary or alternative methods of evaluating such stimuli. These metrics can be recorded while participants view stimuli, without asking them to respond to subjective or objective questions, providing flexibility to researchers in designing experimental studies to measure the quality of these stimuli.

In summary, the results presented above indicate that hypotheses H1 and H3 were supported, hypotheses H2, H4, and H5 were not supported; this result is in agreement with our earlier work [13]. There was a significant difference in the eye movement metrics when participants viewed ASL videos (as compared to when they viewed ASL animations), and some eye movement metrics were significantly correlated with participants' subjective judgments of video and animation quality (grammaticality, understandability, and naturalness of movement).

Specifically, the most notable new findings in this paper are:

- If using proportional fixation time to distinguish between ASL videos and animations, the upper-face AOI should be considered; if using transitions/second, the lower-face AOI should be considered. Since our prior work [13] had not analyzed the eye-tracking data in such a fine-grained manner (i.e., the upper-face and lower-face AOIs had been clumped together into a single "face" AOI), this distinction between them in regard to the significance of transitions per second or proportional fixation time was not identified in that earlier work.
- If seeking an eye metric that correlates with participants' subjective judgments about ASL videos or animations, the time-normalized fixation trail length metric (described in this paper) should be utilized. (The only exception would be for predicting participants' grammaticality judgments for ASL animations: the upper-face proportional fixation time was the best correlated.)

Our lab is studying how to design software that can automatically synthesize ASL animations, and in future work, we will continue to investigate the applications of eye-tracking methodologies in evaluation studies of ASL animations. In current work, we are investigating models of ASL facial expression, and we intend to employ eye-tracking metrics in future evaluation studies. Our goal is to produce understandable ASL animations for deaf people with low English literacy – ultimately leading to better accessibility of educational content for deaf students.

Acknowledgments. This material is based upon work supported by the National Science Foundation under award number 0746556 and 1065009. Pengfei Lu, Jonathan Lamberton, and Miriam Morrow assisted with study preparation and conduct.

References

1. Adamo-Villani, N., Doublestein, J., Martin, Z.: Sign language for K-8 mathematics by 3D interactive animation. Journal of Educational Technology Systems. 33(3), 241-257, (2005)
2. Adamo-Villani, N., Popescu, V., Lestina, J.: A non-expert-user interface for posing signing avatars. Disability and Rehabilitation: Assistive Technology. 8(3), 238-248, (2013)
3. Ardis, S.: ASL Animations supporting literacy development for learners who are deaf. Closing the Gap. 24(5), 1-4, (2006)
4. Cavender, A., Rice, E. A., Wilamowska, K. M.: SignWave: Human Perception of Sign Language Video Quality as Constrained by Mobile Phone Technology. emergency, 7(12), 16-20
5. Cooke, L., and Cuddihy, E.: Using eye tracking to address limitations in think-aloud protocol. In: Professional Communication Conference (IPCC 2005). Proceedings. International pp. 653-658, IEEE, (2005)
6. Duchowski A.: A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments, & Computers. 34, 4, 455-470, (2002)
7. Emmorey, K., Thompson, R., Colvin, R.: Eye gaze during comprehension of American Sign Language by native and beginning signers. J Deaf Stud Deaf Educ. 14, 2, 237-43, (2009)

8. Huenerfauth, M.: Evaluation of a psycholinguistically motivated timing model for animations of American Sign Language. In: Proc. of the 10th international ACM SIGACCESS conference on Computers and accessibility (pp. 129-136). ACM, (2008)
9. Huenerfauth, M.: Spatial and planning models of ASL classifier predicates for machine translation. In 10th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI, (2004)
10. Huenerfauth, M., Lu, P., and Rosenberg, A.: Evaluating importance of facial expression in American Sign Language and pidgin signed English animations. In: Proc. of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (pp. 99-106). ACM, (2011)
11. Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J.: Evaluating American Sign Language generation through the participation of native ASL signers. In: Proc. of the 9th International ACM SIGACCESS Conference on Computers and Accessibility (pp. 211-218). ACM, (2007)
12. Jacob, R. J. K. and Karn, K. S.: Eye Tracking in Human- Computer Interaction and Usability Research: Ready to Deliver the Promises. The Mind's Eye (First Edition). In: Hyönä, J., Radach, R., and Deubel, H. (eds.). Amsterdam: 573-605, (2003)
13. Kacorri, H., Harper, A., and Huenerfauth, M.: Comparing native signers' perception of American Sign Language animations and videos via eye tracking. In: Proc. of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (p. 9). ACM, (2013)
14. Kacorri, H., Lu, P., and Huenerfauth, M.: Evaluating facial expressions in American Sign Language animations for accessible online information. In Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion (pp. 510-519). Springer Berlin Heidelberg, (2013)
15. Karpouzis, K., Caridakis, G., Fotinea, S. E., Efthimiou, E.: Educational resources and implementation of a Greek sign language synthesis architecture. Computers & Education, 49(1), 54–74, (2007)
16. Lu, P., and Huenerfauth, M.: Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. In: Proc. of the 11th international ACM SIGACCESS Conference on Computers and Accessibility (pp. 83-90). ACM, (2009)
17. Mitchell, R., Young, T., Bachleda, B., and Karchmer, M.: How many people use ASL in the United States? Why estimates need updating. Sign Lang Studies, 6(3): 306-335, (2006)
18. Muir, L. J., and Richardson, I. E.: Perception of sign language and its application to visual communications for deaf people. J Deaf Stud Deaf Educ 10, 4, 390-401, (2005)
19. Rayner, K: Eye movements in reading and information processing: 20 years of research. *Psychol Bull* 124(3): 372- 422, (1998)
20. Traxler, C.: The Stanford achievement test, 9th edition: national norming and performance standards for deaf & hard-of-hearing students. J Deaf Stud & Deaf Educ, 5:4, pp. 337-348, (2000)
21. Wagner, M., Marder, C., Blackorby, J., Cameto, R., Newman, L., Levine, P., et al.: The achievements of 100 youth with disabilities during secondary school: A report from 101 the National Longitudinal Transition Study-2 (NLTS2). Menlo 102 Park, CA: SRI International, (2003)