2019
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

Evaluating Sign Language Animation through Models of Eye Movements

Abhishek Suhas Mhatre, Sedeeq Al-khazraji, Matt Huenerfauth

Rochester Institute of Technology, Golisano College of Computing and Information Sciences

asm1610@rit.edu, sha6709@rit.edu, matt.huenerfauth@rit.edu

Abstract

Based on machine-learning models of the eye movements of people watching sign language animations, we predict subjective evaluations of the animation quality.

Keywords

Deaf and Hard of Hearing, Emerging Assistive Technologies, Research & Development

Journal on Technology and Persons with Disabilities
Santiago, J. (Eds): CSUN Assistive Technology Conference

Introduction

There are over 500,000 people in the United States who use American Sign Language (ASL) as a primary form of communication (Mitchell et al., 2006), and many of these individuals prefer to have information provided in ASL, rather than written English. Since providing videos of human signers on websites can lead to significant expense when videos must be re-recorded (when information content is updated), there has been research on methods for automatically producing computer animations of ASL for these users, based on some input script of the desired message (Kacorri et al., 2013). However, to drive progress in the field, researchers need methods for evaluating the quality of the ASL animations that they produce with their software. Unfortunately, designing questions to measure people's comprehension of animations is very time consuming (Kacorri et al., 2014).

For this reason, we investigate whether machine learning algorithms can be trained on eye-tracking data from people who watch ASL animations, to predict whether the person watching the animation judges it to be of high-quality or easy to understand. As discussed in (Huenerfauth and Kacorri 2016), the advantage of this approach is that researchers do not need to design comprehension questions specifically tailored to the information content of the animations shown. Furthermore, by analyzing eye-movements rather than asking overt questions, researchers can avoid artificially drawing participants' attention to specific aspects of the animation, e.g. with questions about particular facial expressions, which could change how the participant views the animation.

Discussion

Problem Statement

In this paper, we present an analysis of prior literature on using machine learning models of eye-movement metrics. Next, we present the results of a study to analyze eye movement recordings of people who are deaf watching computer-generated ASL animations, to determine whether there are patterns in eye-movement that can be used to automatically evaluate the quality of the animations displayed. Using only eye-tracker data as input features to machine learning methods, we present models that can predict whether a user found animations easy to understand and perceived them to be high quality.

Literature Review

Several prior research projects have explored using eye-movement information recorded from humans as input to a machine-learning model, to predict something about the users or about the task being conducted, e.g. for predicting which students in an online learning system will succeed on quizzes (Harper 2015), predicting students' confidence in their answers to test questions (Nakayama and Takahasi 2008), predicting regions of photos that are most salient (Judd et al., 2009), or automatically classifying the genre of documents being read (Kunze et al., 2013).

Prior research has also examined the relationship between the eye-movements of humans viewing sign-language animations and the quality of the animation displayed. In (Kacorri et al., 2013), researchers at our laboratory conducted a study in which participants (Deaf ASL signers) viewed short ASL stories of three types: a video of ASL signer, an animation of ASL with high-quality facial expressions, and an animation of ASL with a no facial expression. After viewing each story, the participant had to answer subjective questions about the quality of the video and

comprehensive questions. Participants more often gazed at the face of an animated ASL signer when the animation was of high quality, and people were more likely to move their eyes between the face and hands of the ASL signer when the animation was of lower quality. In a subsequent study (Kacorri et al., 2014), we found that considering the upper-face and lower-face of the signer as separate "areas of interest" (AOIs) for eye-tracking analysis was valuable. Also, a time-normalized fixation trail length metric was valuable in revealing differences in animation quality.

In the most closely-related prior work to our current study, Huenerfauth and Kacorri (2016) identified metrics that relate to participants' subjective assessment of ASL animations, and they trained a linear regression model on these features. The model correlated with participants' subjective impressions of animation quality, but a limitation of that prior study was that only a single type of model was considered. In this new research, we have explored a wider variety of machine-learning models and eye-movement features, to identify a model that is able to predict participants' assessment of the animations.

Collection of Training Data for this Study

In earlier work from our laboratory (Kacorri et al., 2013), participants (17 Deaf ASL signers) viewed short ASL stories of three types: a video of ASL signer, an animation of ASL with high-quality facial expressions, and an animation of ASL with a no facial expression (Kacorri et al., 2013). After viewing each story, the participant responded to a 1-to-10 scalar question to rate the quality of the ASL animation they had seen: *Grammar:* "Is it grammatically correct?," *Understand:* "Is it easy to understand?," and *Natural:* "Does it move naturally?" In this current study, we utilize the data obtained from these participants as training data for a machine learning task. To structure our modeling as a classification task, we created Boolean target variables *isGrammarHigh, isUnderstandHigh,* and *isNaturalHigh,* based on threshold

values (7, 7, and 5, respectively), which were selected based on an examination of a histogram of the responses for each item, to determine a natural boundary for each. For instance, if a participant had a *Grammar* score of 7 or higher, then we set the value of *isGrammarHigh* to 1, else 0.

Input Features for Machine-Learning Modeling

While a recording of a human's eye-gaze during the entire time a video is viewed would consist of a large stream of x and y coordinates on the screen over time, such data is not suitable for training input to a machine learning task. As discussed in (Harper 2015), we must first calculate summarizing metrics of the eye-movements during the animation, and these metrics are used as input to the machine-learning modeling. Following the approach of (Harper 2015) and the eye-metrics discussed in our prior research (Kacorri et al., 2013; Kacorri et al., 2014), e.g. the percentage of time that someone looks at the upper face of the signer, we calculated several hundred potential input features for our machine-learning modeling. To select an appropriate subset of these attributes to use as input features for our machine learning process, we took our entire training dataset (from all human participants, P1 to P17) as input for the feature selection (results in Table 1). We applied the CFS Subset Evaluator provided in the Weka toolkit (Witten et al., 2016), which calculates the importance of variables by determining the individual predictive ability of every variable individually (based on their correlation with the target variable you are hoping to predict).

Table 1. Final Selected Features for *isGrammarHigh*, *isUnderstandHigh*, and *isNaturalHigh.*

Modeling Target	**Features Selected for Inclusion in Each Model**
isGrammarHigh	Fixations per second, total area-of-interest (AOI) transitions per second, proportion of fixation time (PFT) spent looking at the Hands, PFT spent looking not at Hands nor Face, fixations per second on Head, fixations per second between quadrants, fixations per second on upper left quadrant, percentage of fixations on lower left quadrant
isUnderstandHigh	Fixations per second, total area-of-interest (AOI) transitions per second, transitions per second between Head and Hands AOI, dwells per second total, dwells per second on Head AOI, fixations per second on upper left quadrant, percentage of fixations on lower right quadrant, transitions per second between quadrants, dwells per second on quadrants, dwells per second on upper left (UL) AOI, percentage of dwells on UL AOI
isNatrualHigh	Proportion of fixation time (PFT) spent looking at the Upper Face, PFT, spent looking not at Hands nor Face, percentage of fixations on Hands, total dwells not on Hands nor Face, percentage of dwells on Hands, PFT on upper right quadrant of video, PFT on lower right quadrant, percentage fixations on upper right quadrant, percentage of fixations on lower right quadrant, percentage of dwells on upper left quadrant, percentage of dwells on lower right quadrant

Modeling and Evaluation

For this study, we compared three models: Given our prior research on modeling eye-metrics using regression, we chose to examine a Logistic Regression model, and for comparison, we also trained a J48 Decision Tree and a Support Vector Machine (SVM). For training and testing, we utilized a form of cross-validation, following a "leave data from one participant out at a time" strategy: For each fold of the cross-validation, the testing set consisted of data from one of the 17 human participants in the original study, with the training dataset composed of the

remaining data from the other 16 participants. This process was repeated 17 times, and the average evaluation scores for the 17 "folds" are presented as the values in Table 2.

Table 2. Accuracy Percentage for Each Model.

Model	Grammatical	Understandable	Moves Naturally
Decision Tree	59.31	46.56	54.90
Logistic Regression	54.90	56.86	49.99
SVM	61.76	56.37	50.98
Baseline (Majority)	66.17	54.90	50.98

Based on the results of this modeling and evaluation process, we can conclude that for at least some aspects of human subjective judgments about the quality of ASL animations (whether it is understandable or whether they believe it moves in a natural manner), we can use patterns in the movements of a human's eye-gaze (when watching the animation) to predict their opinion of the animation quality. On the other hand, for some aspects of human judgments about ASL animations – specifically, whether the animation is grammatically correct – we were not able to predict these judgments well, based solely on the eye-movement patterns of humans watching the animations. (The performance of our best model was well below the baseline level, which simply selected the majority classification in our dataset.)

Conclusions

The results of this study could be used to build an automatic prediction system, to allow researchers to evaluate the quality of an ASL animation by asking participants to watch animations while their eye-movements are recorded; the participants would not need to answer

any questions as part of the evaluation. This additional flexibility in designing evaluation studies of ASL animation technology, which may benefit people who are Deaf or Hard of Hearing.

A limitation of this study was the relatively small dataset of 17 human participants, who each viewed 12 ASL animation videos; in future work, we would like to increase our training dataset by collecting additional eye-movement recordings from humans observing ASL animations. With a larger dataset, we would like to consider additional eye-movement metrics.

Works Cited

Harper, Allen V. R. “Eye Tracking and Performance Evaluation: Automatic Detection of User Outcomes.” 2015. The *City University of New York. PhD dissertation*. https://academicworks.cuny.edu/gc_etds/964/.

Huenerfauth, Matt, and Hernisa Kacorri. “Eyetracking Metrics Related to Subjective Assessments of ASL Animations.” *Journal on Technology and Persons with Disabilities, 2016 California State University,* Northridge, 2016.

Judd, Tilke, et al. “Learning to predict where humans look.” *Computer Vision, 2009 IEEE 12th international conference on, IEEE*, 2009.

Kacorri, Hernisa, et al. “Comparing Native Signers’ Perception of American Sign Language Animations and Videos via Eye Tracking.” *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS ’13), ACM*, New York, NY, USA, 2013. DOI: http://dx.doi.org/10.1145/2513383.

Kacorri, Hernisa, et al. “Measuring the Perception of Facial Expressions in American Sign Language Animations with Eye Tracking.” *In: Stephanidis C., Antona M. (eds) Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice, (UAHCI 2014), Lecture Notes in Computer Science*, vol. 8516, Springer, Cham, 2014.

Kunze, Kai, et al. “I know what you are reading: Recognition of Document Types Using Mobile Eye Tracking.” *Proceedings of the 2013 International Symposium on Wearable Computers. ACM*, Zurich, Switzerland, 2013.

Mitchell, Ross E, et al. “How many people use ASL in the United States?: Why estimates need updating.” *Sign Lang Studies*, vol. 6, no. 3, 2006, pp. 306-335.

Nakayama, Minoru, and Yosiyuki Takahasi. “Estimation of certainty for responses to multiple-choice questionnaires using eye movements.” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 5, no. 2, 2008.

Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. *Morgan Kaufmann*, 2016.