Expression: A Dyadic Conversation Aid using Google Glass for People with Visual Impairments

ASM Iftekhar Anam
Dept. of Electrical & Computer Engineering
University of Memphis
Memphis, TN USA
aanam@memphis.edu

Shahinur Alam
Dept. of Electrical & Computer Engineering
University of Memphis
Memphis, TN USA
salam@memphis.edu

Mohammed Yeasin
Dept. of Electrical & Computer Engineering
University of Memphis
Memphis, TN USA
myeasin@memphis.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
UbiComp'14 Adjunct, September 13–17, 2014, Seattle, WA, USA.
Copyright 2014 978-1-4503-3047-3/14/09...$15.00.
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

Abstract
Limited access to non-verbal cues hinders the dyadic conversation or social interaction of people who are blind or visually impaired. This paper presents **Expression** — an integrated assistive solution using Google Glass. The key function of the system is to enable the user to perceive social signals during a natural face-to-face conversation. Empirical evaluation of the system is presented with qualitative (Likert score: $4.383/5$) and quantitative results (overall F-measure of the nonverbal expression recognition: 0.768).

Author Keywords
Wearable system; Visually Impaired; Assistive Technology; Facial Expression Recognition

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces- *Evaluation/ methodology, User-centered design*; K.4.2 [Computers and Society]: Social Issues- *Assistive technologies for persons with disabilities*

General Terms
Information Interfaces and Presentation, Assistive technologies for persons with disabilities

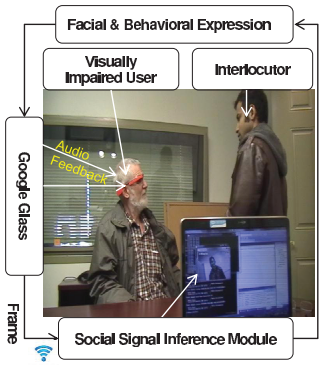

Figure 1: Expression in a social interaction

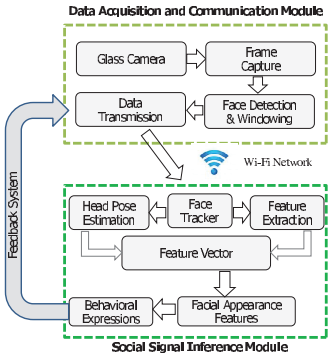

Figure 2: System Diagram of the **Expression**

Introduction

Assistive technology solutions can help people with disability from social isolation, depression, facilitate social interaction and enhance the quality of life.

In this paper, we focus on developing a ubiquitous system to provide access to non-verbal cues in social interaction for people who are blind or visually impaired (henceforward termed as a representative user). The design team had a number of brainstorming sessions with the representative users to determine the key functionalities of the system. It was unanimously agreed that access to the interlocutor's appearance, facial features, and behavioral expressions would significantly help in dyadic conversation.

Towards developing a dyadic conversation aid we present a fully integrated system, called **Expression** using Google Glass It is designed to predict interlocutor's social signals (such as facial appearance features, behavioral expressions, and emotions) and provide real-time feedback to the user in an unobtrusive manner. With **Expression** application installed, it captures video stream ($5-8$ Frames per second) using the Google Glass (henceforth termed as Glass) camera and transmit it to a server. The server analyzes the images to detect facial features and predict behavioral expressions and return them to the Glass. Speech feedback is conveyed to the user using built-in Text-to-Speech service. Both objective and subjective evaluation was performed to illustrate the utility of the system. Subjective evaluation of **Expression** was performed using a five (5) point Likert Scale and was found to be excellent(4.383).

System Design

Design and implementation of **Expression** followed the ideas from participatory design and went through a number of iterations. It consists of three modules (Figure 2): (a) data acquisition and communication; (b) social signal inference engine; and (c) feedback system. Figure 1 shows the **Expression** being used by a participant. The data acquisition module captures video stream using the Glass camera. Initially, a Viola-Jones face detector [7] was used for fast detection of a facial region. The user receives speech feedback about the face position and size on the screen. This helps the user to self-correct her posture or orientation to capture the interlocutor's face at the center of the screen. The Glass then transmits captured frames to a server. The social signal inference engine at the server computes facial features and estimates head pose to generate feature vectors. A rule-based classifier infers the behavioral expressions from feature vectors using a temporal sliding window. The output is then sent to the feedback module to generate speech feedback.

A set of non-verbal behaviors (Figure 3) were selected through participatory design process that are also supported by Psychology literature [1]: (a) head movements (look up/down, look left/right, tile left/right); (b) facial expressions (smile, open smile); and (c) behavioral expressions (yawn, sleepy).

Data Collection and Annotation

Since the interlocutor can be either sighted or have disabilities, both types of people (6 visually impaired and 14 sighted) were included in data collection. Each of them was asked to participate in a 10 minute face-to-face conversation on topics of their interest. Data were recorded using a Glass where one of the designers acted as a user and the subjects as interlocutors. After cleaning up noises from the recordings, five annotators performed frame by frame annotation of the non-verbal behaviors. The inter-rater agreement measured by Fleiss' Kappa was

Figure 3: Example expressions (from top-left): Tilt Left,Looking Right, Looking Left, Yawn, Smile, Looking Up, Looking Down, Tilt Right

0.791 which indicates a significant agreement.

Feature Selection and Modeling
Facial features and head pose are extracted using a Constrained Local Model (CLM) based face-tracker developed by Saragih [5] et al. It fits a parameterized 3D shape model and track 66 facial landmark points in the image. We use distance based features such as (a) height of inner eyebrow and outer eyebrow; (b) height of eye-opening; (c) height of inner and outer lip boundary; and (d) distance between lip corners. A canonical reference shape is obtained from the mean of the shapes of different expressions. The features are calculated as the ratio of the distances between appropriate landmarks after removal of global transformation. The head pose is estimated from the tracked 3D shape. The rules of the classifier proposed in [4] were formulated using thresholds learned from the annotated data.

Development through Participatory Design
We worked in a participatory design process with a focus group comprised of three visually impaired individuals recruited through Clovernook Center for the Blind and Visually Impaired and the Mid-South Access Center for Technology (Mid-South ACT).

Due to limited battery life, the Glass is not suitable for continuous use in a mobile application. Moreover, the device gets overheated if the camera is used continuously. Likamawa et al [2] investigated various use cases to quantify the power consumption and characterize temperature profile of the Glass. The alpha prototype of **Expression** was designed only to capture and send frames to a server. That caused some lag in finding a face when the application starts. To address it, a Viola-Jones face detector was incorporated to the application. However, detecting faces in every frame caused a drop in frame rate and the device gets overheated quickly. Therefore, the face detector was used only at the start of the application to find a face and after receiving a message of tracking failure. We also provide feedback to the user to move closer towards the interlocutor when the face size is smaller than 40×40 on the screen. We empirically set the frame rate $5 - 8$ FPS to ensure reliable expression recognition and reduced overheating.

Feedback Design
Based on a survey with representative users it was concluded that speech feedback through ear bud is the preferred mode for the **Expression**. This does not require any training as compared to sensory substitution systems such as vOICE [3] and iFEPS [6]. In the initial prototype, continuous feedback of the spotted behavioral expressions was provided to the user. Based on the suggestions from participatory design team, the feedback is generated only when there is change in expressions to reduce distraction. The participants reported that it was easy to follow the conversation with the discrete feedback.

Evaluation Results and Discussion

Both objective and subjective evaluation (with ten subjects) were performed to illustrate the utility of the system. Table 1 shows the quantitative analysis of **Expression** system. The overall F-measure is 0.768 which is reasonable with natural expressions data. It is conspicuous to note that the system spots a lot of false positives. It can be attributed to overlapping expressions and involuntary movements. The participants were asked to complete a set of usability questionnaire using a 5-point Likert scale (5 being the highest). Figure 4 shows the qualitative evaluation result. The comparatively low score of "Willing to Use" can be explained by the perceived uncertainty in social acceptance of a new device such as the Glass.

Expressions	Precision	Recall	F-Measure
Smile	0.913	0.778	0.840
OpenSmile	0.833	0.833	0.833
Sleepy	0.75	0.692	0.720
Yawn	0.625	0.714	0.667
Looking up/down	0.895	0.708	0.791
Looking left/right	0.933	0.636	0756
Average	**0.825**	**0.727**	**0.768**

Table 1: Performance of the Social Signal Inference Engine

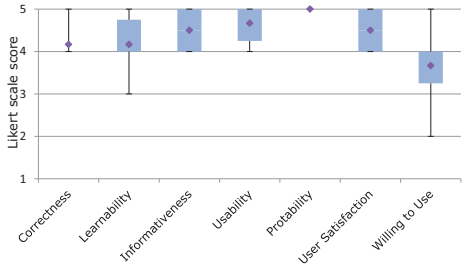

Figure 4: Box plot of Qualitative Evaluation of the **Expression**

As future work, we will experiment with multi-modal feedback since one visually impaired participant wanted tones along with the speech feedback. According to her if she misses the speech feedback while concentrating on the conversation, tones will be helpful. We plan to increase the number of behavioral expressions and use machine learning technique to improve the model. Moreover, we shall utilize sensor data to distinguish the movements of the user and the interlocutor, and conduct long term studies with the **Expression**.

Acknowledgments

We are thankful to Dr. Lavonnie Claybon, Director of Midsouth ACT to provide access to subjects to conduct the study.This work was partially funded by National Science Foundation (NSF–IIS-0746790), USA. Any opinions, findings, and conclusions or recommendations do not reflect the views of the funding institution.

References

[1] Argyle, M., Alkema, F., and Gilmour, R. The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology 1*, 3 (1971), 385–402.

[2] LiKamWa, R., Wang, Z., Carroll, A., Lin, F. X., and Zhong, L. Draining our glass: An energy and heat characterization of google glass. *arXiv preprint arXiv:1404.1320* (2014).

[3] Meijer, P. B. An experimental system for auditory image representations. *Biomedical Engineering, IEEE Transactions on 39*, 2 (1992), 112–121.

[4] Rahman, A., Anam, A. I., Tanveer, M. I., Ghosh, S., and Yeasin, M. Emoassist: A real-time social interaction tool to assist the visually impaired. In *15th International Conference on Human-Computer Interaction (HCII 2013)* (2013).

[5] Saragih, J. M., Lucey, S., and Cohn, J. F. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE (2009), 1034–1041.

[6] Tanveer, M. I., Anam, A., Yeasin, M., and Khan, M. Do you see what i see?: designing a sensory substitution device to access non-verbal modes of communication. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ACM (2013), 10.

[7] Viola, P., and Jones, M. J. Robust real-time face detection. *International journal of computer vision 57*, 2 (2004), 137–154.