Mobile Dual Eye-Tracking Methods: Challenges and Opportunities

Alan T. Clark[*] and Darren Gergle[*†]
[*]Department of Communication Studies
[†]Department of Electrical Engineering and Computer Science
Northwestern University
{alan-clark@northwestern.edu, dgergle@northwestern.edu}

Abstract. Mobile dual eye-tracking approaches allow researchers to develop more nuanced and ecologically valid accounts of how interlocutors coordinate their gaze during natural conversation. This approach can be particularly fruitful for studying multimodal reference by allowing researchers to incorporate the effects of physical movement and spatial positioning of speakers and listeners into their models. We discuss the research opportunities we envision and methodological challenges we have encountered as part of our mobile dual eye-tracking approach.

Background

Dual eye-tracking methods have provided great opportunities for researchers to better understand the role of gaze as a coordination mechanism in conversation. One particularly fruitful avenue of research examines the role gaze plays in reference. Reference is how people specify the person, object, or entity that they are talking about (Carlson, 2004). Traditional accounts of reference have focused on the important role spoken language plays, yet reference is often a multimodal process, with objects being evoked through a speaker's actions, movement, or other pragmatic contextual cues such as gestures or head nods. These are all part of interlocutors' toolkit of 'conversational resources' (Kraut, Fussell, & Siegel,

2003) which they can use to successfully ground their references.

Several recent studies have captured how the dynamics of gaze between speaker and listener influence the production and interpretation of referential descriptions. Hanna and Brennan (Hanna & Brennan, 2007) found that addressees use the speakers' gaze as a cue for disambiguating references, often before the reference would be disambiguated linguistically. Nakano and colleagues (Nakano, Reinstein, Stocky, & Cassell, 2003) found that speakers look at their addressees in order to ground references to new entities. Bard and colleagues (Bard, Hill, & Arai, 2009) found that, in a computer-mediated task, partners' rates of shared gaze towards objects was not high, but well above chance. By tracking speakers' and listeners' gaze as they talk about their environment and the objects in it, we have been able to develop a more complete picture of reference behavior in conversation and collaboration.

However, most prior accounts that study reference using eye-tracking methods take place in controlled contexts with static referential domains, with interlocutors typically seated and immobile. Yet in everyday conversation, people move around, shift their body positions, and gaze at different locations around their shared space (Hindmarsh & Heath, 2000). Reference behaviors, and the dynamics of gaze, do not take place in the static contexts often favored by researchers. Interlocutors' mobility can have a considerable effect not only on rates of shared gaze, but on how interlocutors use gaze as a conversational resource to support reference. For example, our prior work (Gergle & Clark, 2011) found that mobile pairs' rate of gaze overlap on objects they are referring to is generally high, but actually occurs at below-chance levels when talking about nearby objects with local deictic forms ("this", "these"). In such situations, speakers often spatially evoke the object through their movement, making the object salient and their attention to it implicit, which seemingly marginalizes gaze as means for evoking the referential object.

Findings such as these highlight the potential importance of using a *mobile* dual eye-tracking method to explore reference in more natural conversational contexts. In the following sections, we describe the novel system and methods developed in our prior work (Clark & Gergle, 2010; Gergle & Clark, 2011) to support studies of mobile dual eye-tracking, discuss technical and methodological challenges with such an approach, and discuss the opportunities for future work it affords. In doing so, we aim to prepare other researchers for using mobile dual eye-tracking methods in a way that preserves ecological validity without excessive compromise in data quality and experimental control.

Our Approach

In our mobile dual eye-tracking research (Clark & Gergle, 2010; Gergle & Clark, 2011), we sought to study how multimodal reference, and particularly the interactions of gaze coordination and language use, differed in mobile and stationary contexts. Pairs were randomly assigned to either of two seated conditions (across a table; side-by-side) or one standing mobile condition. We employed a naturalistic conversation task, in which participant pairs had to collaboratively rank LEGO objects on the basis of their likelihood to be replicas of modern art. This was primarily a conversation elicitation task, designed to stimulate discussion, as we were concerned that zero-history experimental pairs are often reticent at best. This worked well, with pairs talking for an average of over five and a half minutes and producing an average 50.3 references, with no significant differences across conditions. Although we do not plan to use tasks designed exclusively for reference production in the future, we still plan to employ the conversation elicitation approach because of its success in producing a sizable corpus of references.

In order to capture the participant pairs' gaze patterns as they talked about the LEGO "art", we used a combination of off-the-shelf equipment with custom in-house software to facilitate our process. We used a pair of Applied Science Laboratories (ASL) MobileEye MKII eye-trackers to record participant pairs' gaze. The MobileEye trackers recorded directly onto a digital video recorder, which participants would carry around in a small waist pack. The cable between the trackers and the recorder was the only tether our participants had. This allowed our participants full motion of their head and upper body, although we asked them not to pull on the cables or to touch the eye-trackers themselves.

Using the point-of-gaze data and scene video from the trackers, we were able to automatically detect when they were looking at objects in their environment. We developed custom software that makes use of the ARToolkit 2.7 (Kato & Billinghurst, 1999) and permits the automatic capture and comparison of point-of-gaze data to the location of actual objects in 3D space. The architecture of this system is shown in Figure 1.

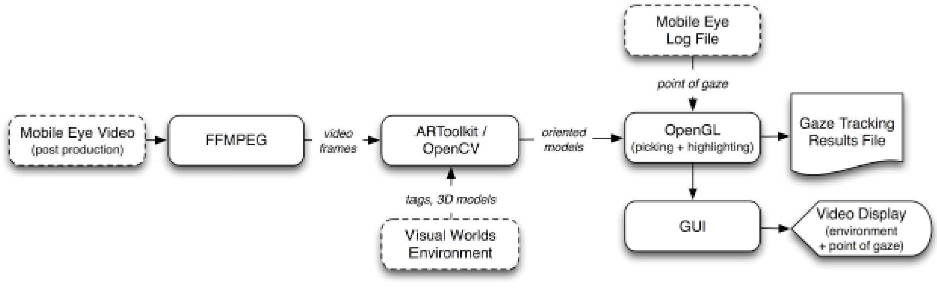
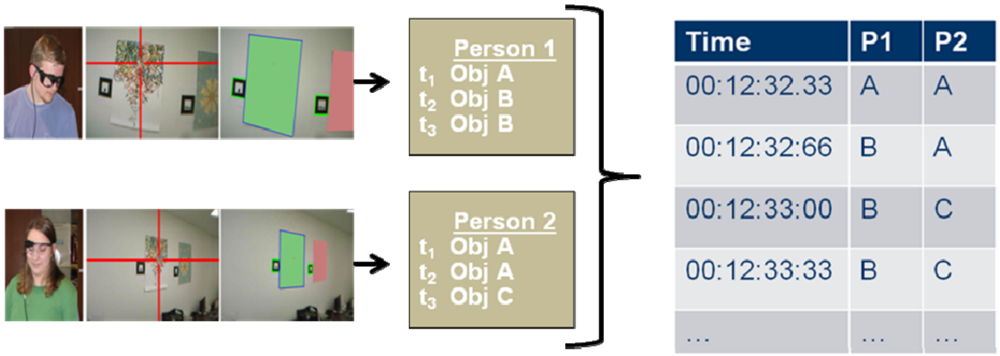

Mobile Eye Log File
point of gaze
Mobile Eye Video (post production)
FFMPEG
video frames
ARToolkit / OpenCV
oriented models
OpenGL (picking + highlighting)
Gaze Tracking Results File
tags, 3D models
Visual Worlds Environment
GUI
Video Display (environment + point of gaze)

Figure 1. Architecture of augmented reality based automated gaze recognition system

By placing vision markers next to the LEGO objects we employed as stimuli, we were able to track their location in the environment as participants looked and moved around their space. When the markers and associated objects appeared in the scene video, we could automatically look at both participants' point of gaze at a given moment to determine whether their gaze was falling on a given object (see Figure 2). This saved an extraordinary amount of time in behavioral coding and helped to overcome logistical concerns (e.g. budget for undergraduate coders) that might undermine plans for mobile dual eye-tracking studies. We note that some current systems are able to achieve similar outcomes using infrared markers placed in the environment, although our approach may be more flexible for studies where stimulus objects have complex shapes or will be viewed from multiple directions.



Time	P1	P2
00:12:32.33	A	A
00:12:32:66	B	A
00:12:33:00	B	C
00:12:33:33	B	C
…	…	…

Figure 2. Output from our automated gaze recognition system

To analyze our gaze and linguistic data, we used generalized linear mixed-models regression with covariance modeling to account for the fact that dyadic gaze data is non-independent (Barr, 2008). This method is available in standard statistical packages including STATA and R. To operationalize gaze coordination, we used the proportions of overlapping points-of-gaze on referent objects (sampled at 30hz) in a window surrounding the onset of the referring expression. We also linguistically coded each reference for a variety of linguistic variables, including which object was the speaker's intended referent – that is, which of the objects in their environment they were trying to pick out. Despite some initial concerns, our coding of speaker's intended referent was reliable (Cohen's kappa = .80). Although this coding was time-intensive, it allowed us to look at gaze overlap in a way that was sensitive to the pair's discourse context at a given moment. For example, we could look at situations where the addressee was not looking at the speaker's intended referent before they named it, to see if the speakers tended to use more demonstrative or detailed language. We were also able to use this approach to show above-chance levels of gaze coordination to the speaker's intended referent. Generally, these methods allowed us to paint a more sophisticated picture of how gaze was used by speaker and listener, moment-to-moment, than less context-sensitive approaches (e.g. cross-recurrence analysis).

Challenges

Although we piloted and extensively pre-tested our dual eye-tracking methods, we identified a number of methodological issues, some of which we anticipated and some we encountered on the way. In this section, we relate some challenges that could affect other researchers using a similar approach. We discuss issues with experiment setup, calibration, and accounting for added risks of lost data.

First, mobile dual eye-tracking can face subtle but important problems with equipment, setting, and stimuli that need to be accounted for. Lighting can vary from place to place even within a controlled experiment room, potentially creating glare on eye-trackers that use mirrors that can distract participants. Inconsistent lighting can also be a problem for automated processing of video, such as our automated gaze coding system. Positioning and operation of cameras in the room to record movements, gestures, and other non-verbal visible behaviors requires additional expertise beyond just “point and shoot”, at risk of losing data if mobile participants occupy blind spots. Movement around space can affect the salience of stimuli – for example, if one were tracking gaze toward sculptures as participants walked through a museum, one’s results might be confounded if some participants could view sculptures from all angles while others could only view sculptures from one direction.

One also needs to carefully consider how to record and code movement, as different approaches support different research goals. We were interested in high-level differences between users who could move during conversation and users who could not, and (currently) effects of relative positioning between speaker, listener, and intended referent. For us, a spatial coding grid system and cameras in the room to record users’ positions around a table were sufficient. In outside-the-lab contexts, or in studies that want to identify effects of specific types or timing of movement, technological solutions such as head-tracking are likely required.

Proper calibration, which is essential to getting high-quality data, can be more challenging for mobile than stationary dual eye-tracking research. One has to be careful about calibrating to make sure the vertical angle on the head-mounted scene cameras aligns it with where the participant will typically be looking. For example, if stimulus objects are on a waist-height table, but one calibrates with eye-level objects on the wall, the objects of interest are likely to be partially occluded or entirely out of frame. For similar reasons, it is useful to calibrate participants in the stance (i.e. seated vs. standing) that they will primarily occupy during the experiment. This can also have added benefits for calibrating the tracker to the participant’s eye, as their viewing angle toward the objects can affect where one needs to position mirrors or infrared lenses to get a clear, unobstructed view of the pupil.

Finally, it is particularly important in mobile dual eye-tracking studies to oversample and run more participants than seemingly needed. In all dual eye-tracking research, there is a fairly high chance that data will have to be thrown out if one participant's data is unusable due to improper calibration, physical features (e.g., long eyelashes) that impede proper tracking, and so forth. The risk of losing data is slightly increased in mobile dual eye-tracking, an unfortunate byproduct of giving participants the freedom to move around. For example, if cables are not properly affixed and close to a participant's body it can catch on things as they pass; however, if the cable from the recording unit to the eye-tracker does not have enough slack, it can become detached. Researchers using mobile dual eye-tracking need to be aware of this increased risk and plan for more participants than the minimum needed to acquire the desired statistical power.

Opportunities

Despite technical challenges with mobile dual eye-tracking, its ability to provide rich gaze coordination data in mobile contexts opens promising avenues for research. Mobile dual eye-tracking allows us to study dynamics of gaze in more natural, fluid, ecologically valid contexts. In everyday conversations, people and their referential domains shift as they move and reposition themselves through their shared space. Even in stationary contexts, people often shift their posture, reposition their chairs, or look over each other's shoulders. For non-mobile eye trackers, capturing this type of interaction ranges from impractical to impossible.

Exploring gaze in these more natural contexts can also reduce the disparity between experimental findings and the real-world situations addressed in the standard "implications for design" section. Similarly, it can allow us to specifically explore how gaze coordination might be used as a form of input in mobile, ubiquitous computing, or otherwise hands-free technological systems. For example, one might design a system that uses the gaze dynamics of collocated pairs (e.g. a repair crew) to capture their attention and focus, and to direct cameras for a remote observer or team member (e.g. an engineer).

Finally, using mobile dual eye-tracking methods allows us to refine our understanding of findings from static referential contexts, building a more sophisticated account of the role of gaze coordination. As mentioned above, we found that pairs have low rates of gaze coordination when talking about spatially proximal objects, often because speakers would 'spatially evoke' them with movement. This showed that the role of gaze coordination in reference is flexible and situational. Going forward, researchers might use mobile dual eye-tracking to add yet more nuance to our practical and theoretical understanding of gaze coordination, looking at a variety of spatial contexts and types of mobility. We only looked at mobility for users moving around a room with objects on a table in the middle, and the effect of mobility on gaze coordination might be rather different for students touring a museum or shoppers browsing the aisles. Our next step is to develop predictive models of how the relative position of speaker, listener, and objects might be used to predict the speaker's intended referent.

Conclusion

Research using mobile dual eye-tracking methods is important for developing accounts of cognitive and social effects of gaze coordination in a range of collaborative contexts. We argue that mobile dual eye-tracking methods are particularly useful for developing ecologically valid accounts that can be readily applied in real-world technologies. We encourage other researchers to use mobile dual eye-tracking approaches, but in doing so to be prepared for the methodological challenges that come with studying the gaze of mobile users.

Acknowledgements

We would like to thank Isaac Wilson and Steve Howard. Funding provided by National Science Foundation grant #0953943, and an NSF GRFP award.

References

Bard, E. G., Hill, R., & Arai, M. (2009). Referring and gaze alignment: Accessibility is alive and well in situated dialogue. In *Proceedings of the Cognitive Science Society*.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory & Language, 59*, 457-474.

Carlson, G. (2004). Reference. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 74-96): Blackwell Publishing Ltd.

Clark, A., & Gergle, D. (2010). Effects of Shifting Spatial Context on Referential Form. In *Proceedings of the 20th Annual Meeting of the Society for Text and Discourse*.

Gergle, D., & Clark, A. (2011). See What I'm Saying? Using Dyadic Mobile Eye Tracking to Study Collaborative Reference. In *Proceedings of CSCW 2011*, pp. 435-444.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596-615.

Hindmarsh, J., & Heath, C. (2000). Embodied reference: A study of deixis in workplace interaction. *Journal of Pragmatics*, *32*, 1855-1878.

Kato, H., & Billinghurst, M. (1999). Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System. In *Proceedings of IWAR 99*.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human Computer Interaction*, *18*(1), 13-49.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Association for Computational Linguistics*, 553-561.